University of California
Santa Barbara

Quantum Chaos, Operator Growth, and Holography

A dissertation submitted in partial satisfaction
of the requirements for the degree

Doctor of Philosophy
in
Physics

by

Alexandre Albert Streicher

Committee in charge:

Professor Joesph Polchinski, Chair, *in memoriam*
Professor Donald Marolf, Chair
Professor Nathaniel Craig
Professor Dirk Bouwmeester

June 2019

The Dissertation of Alexandre Albert Streicher is approved.

Professor Nathaniel Craig

Professor Dirk Bouwmeester

Professor Donald Marolf, Committee Chair

June 2019

Dedicated to Joesph Polchinski

Acknowledgements

This is the most difficult section of the thesis to write, as I can only put into words my feelings of thanks towards everyone who has helped me along the way.

First, I would like to thank my mom and dad, the people from whom I could always feel strong feelings of love and pride no matter what. We're not a usual family, but I still feel happy.

Next, I want to thank everyone who helped me when I was feeling down, without whom I would not have achieved this thesis. During my early years at UCSB, I was often comforted during dark times by Jason, Joe, and Nate. At Stanford, Phil, Steve, Eva, and Lenny took lots of time to talk to me. Then, I found myself supported by many people, such as Guy, Dan, Edgar, Seth, Tarek, Adam, Zhenbin, and Ying. Furthermore, countless others were extremely kind to me such as Matt, Chaitanya, Richard, Brandon, Geoff, Dan, Ahmed, Aron, and Mae. Although it has been somewhat pressuring to have so many people believing in me, it has also made me feel warm inside.

Academically, I must thank Don, Joe, Steve, and Lenny for investing time and resources into me. They all took a chance on me, and I'm happy to say that something nice has come out of it.

Curriculum Vitæ

Alexandre Albert Streicher

Education

2019	Ph.D. in Physics (Expected), University of California, Santa Barbara.
2017	M.A. in Physics, University of California, Santa Barbara.
2014	B.A. in Physics, Cornell University

Publications

1. A. Streicher, "Wringing Four-Point Functions from Euclidean-Twisted Two-Point Functions,". To appear late Jun 2019.
2. X.-L. Qi and A. Streicher, "Quantum Epidemiology: Operator Growth, Thermal Effects, and SYK," arXiv:1810.11958 [hep-th].
3. A. R. Brown, H. Gharibyan, A. Streicher, L. Susskind, L. Thorlacius, and Y. Zhao, "Falling Toward Charged Black Holes," arXiv:1804.04156 [hep-th].
4. D. A. Roberts, D. Stanford, and A. Streicher, "Operator growth in the SYK model," *JHEP* **06** (2018) 122, arXiv:1802.02633 [hep-th].
5. J. S. Cotler, G. Gur-Ari, M. Hanada, J. Polchinski, P. Saad, S. H. Shenker, D. Stanford, A. Streicher, and M. Tezuka, "Black Holes and Random Matrices," *JHEP* **05** (2017) 118, arXiv:1611.04650 [hep-th].

Abstract

Quantum Chaos, Operator Growth, and Holography

by

Alexandre Albert Streicher

The exact role of the internal degrees of freedom (a.k.a. d.o.f.) in holography is not well-understood. Thus, in this thesis, we study a toy model of holography without space: the Sachdev-Ye-Kitaev (SYK) Model. This 0+1 theory of all possible 4-body interactions of N fermion "flavors"/"colors" features a low energy limit reproducing aspects of 1+1 Jackiw-Teitelboim gravity. First, we show that the inherent discreteness of the quantum spectrum results in universal late-time behavior due to eigenvalue repulsion. We then note that the theory's four-point functions probe the phenomenon of operator growth, where an internal d.o.f. goes on to epidemically evolve into larger products of internal d.o.f.s. In this manner where small operators smoothly grow into superpositions of increasing products of operators, we observe a sort of "size" locality, which is intimately tied with the notion of a conformal primary "descending" along its descendants. In fact, we find that the underlying structure of the SYK epidemic limits to that of a probe particle falling into a AdS_2 black hole. In other words, similar to how nearest neighbor interactions lead to dynamics on a flat space background, we demonstrate that many internal interactions lead to dynamics on a higher dimensional geometry.

Contents

Chapter 1

Introduction

In a sense, there is much we do not understand about string theory, as certain problems one would hope to solve with such a complete theory of quantum gravity become either impossibly complicated or completely unapproachable. Thus, it is important to learn more about the various mechanisms within string theory while also attempting to distill universal aspects of quantum gravity. One example is the holographic principle, originally motivated by the large but area-scaling nature of black hole entropy. Its manifestation in string theory arises due to open-closed string duality, which allows one to equate certain string theories and lower-dimensional quantum field theories with many internal degrees of freedom. As a result, various gravitational phenomena can be thought of as reflecting particular aspects of quantum chaos.

To isolate the role of the many internal degrees of freedom in holography, we shall turn our attention to a toy model with no spatial degrees of freedom. One takes N flavors of quantum mechanical Majorana fermions, couples them together in all possible q-body interactions, and examines the low-energy limit. Known as the Sachdev-Ye-Kitaev (SYK) model, one then finds semi-classical equivalence

with Jackiw–Teitelboim gravity, a $1+1$-dimensional theory that effectively describes the decoupled thermal atmosphere of near-extremal Reissner-Nordstrom black holes [1–4].

Quantum chaotic features of the many-body theory manifest non-trivially in the holographic dual. In chapter 2 we explore the very late-time phenomena of eigenvalue repulsion in the SYK model and comment on its quantum gravitational implications. Turning our attention to earlier times, we then examine the behavior of many-body four-point functions, which are known to obey a universal bound saturated by gravity [5]. We note that fermionic four-point functions directly measure the operator growth of simple fermions into larger products of fermions at the same point. In chapter 3, we explicitly determine this operator distribution at infinite temperature for a Heisenberg-evolved fermion in the SYK model and find that a rich structure emerges in the $N \gg q \gg 1$ limit. In chapter 4, we develop the technology necessary to understand and calculate this structure at arbitrary temperature and find many fascinating features. Most interestingly, after a thermal timescale $t \sim \beta$, the operator distribution of the fermion $\psi_1(t)$ along larger products of fermions exactly matches the momentum wavefunction of a particle falling towards the horizon of a AdS_2 black hole.

1.1 Permissions and Attributions

1. The content of chapter 2 and appendix A is the result of a collaboration with Jordan Cotler, Guy Gur-Ari, Masanori Hanada, Joseph Polchinski, Phil Saad, Stephen Shenker, and Douglas Stanford. It has previously appeared

in the Journal of High Energy Physics [6].

2. The content of chapter 3 is the result of a collaboration with Daniel Roberts and Douglas Stanford. It has previously appeared in the Journal of High Energy Physics [7].

3. The content of chapter 4 is the result of a collaboration with Xiao-Liang Qi. It will soon appear in the Journal of High Energy Physics [8].

Chapter 2

Black Holes and Random Matrices

We argue that the late time behavior of horizon fluctuations in large anti-de Sitter (AdS) black holes is governed by the random matrix dynamics characteristic of quantum chaotic systems. Our main tool is the Sachdev-Ye-Kitaev (SYK) model, which we use as a simple model of a black hole. We use an analytically continued partition function $|Z(\beta+it)|^2$ as well as correlation functions as diagnostics. Using numerical techniques we establish random matrix behavior at late times. We determine the early time behavior exactly in a double scaling limit, giving us a plausible estimate for the crossover time to random matrix behavior. We use these ideas to formulate a conjecture about general large AdS black holes, like those dual to 4D super-Yang-Mills theory, giving a provisional estimate of the crossover time. We make some preliminary comments about challenges to understanding the late time dynamics from a bulk point of view.

2.1 Introduction

One of the deep questions in quantum gravity is the origin of the discrete spectrum of black hole microstates, from the *bulk* perspective of holographic duality. For large black holes the AdS/CFT duality makes the answer clear from the boundary perspective — a boundary field theory on a compact space generically has a discrete spectrum of states. But its origin from bulk gravity or string theory, even including nonperturbative effects like branes, is still mysterious.

Maldacena [9] pointed out a signature of a discrete energy spectrum that can (in principle) be computed in the bulk — the lack of decay of two-point functions evaluated at very late time. Dyson, Lindesay, and Susskind [10] applied these ideas to the study of correlators in de Sitter space.

To understand the way in which a two-point function diagnoses a discrete energy spectrum we can express it in the energy basis. The two-point correlation function of a Hermitian operator[1] $O(t)$ at inverse temperature β is given by

$$\begin{aligned} G(t) &= \frac{1}{Z(\beta)} \operatorname{tr}\left[e^{-\beta H} O(t) O(0)\right] \\ &= \frac{1}{Z(\beta)} \sum_{m,n} e^{-\beta E_m} |\langle m|O|n\rangle|^2 e^{i(E_m - E_n)t} \,. \end{aligned} \tag{2.1}$$

Here, $Z(\beta) = \operatorname{tr}\left(e^{-\beta H}\right)$ is the partition function and $|n\rangle$ are energy eigenstates with energies E_n. At early times we can replace the sum over eigenvalues by a coarse grained integral over a smooth density. $G(t)$ will generically decay exponentially in time, but the decay does not continue indefinitely. At late times the

[1] We assume that in a quantum field theory the operator is suitably smeared to eliminate any short distance divergences.

discreteness of the spectrum becomes important, and the phases in (2.1) cause $G(t)$ to oscillate rapidly and erratically. The correlation function is exponentially small and no longer decays.

Holographically the coarse grained approximation is equivalent to a perturbative gravity calculation, and the exponential decay to quasinormal mode behavior [11]. The decay continues forever in this approximation.

There is a somewhat simpler diagnostic of a discrete energy spectrum, introduced in the black hole context by [12]. We define

$$Z(\beta, t) \equiv \text{tr}\left(e^{-\beta H - iHt}\right) . \tag{2.2}$$

The quantity $Z(\beta, t)$ can be obtained by starting with $Z(\beta)$ and analytically continuing $\beta \to \beta + it$. At late times $Z(\beta, t)$ also oscillates erratically.

The time average of an observable and its moments is a simple way to quantify its late time behavior.[2] In fact, the time average of $Z(\beta, t)$ vanishes, which means that at late times this observable fluctuates around zero. The typical size of the fluctuations can be studied by considering the squared quantity

$$\left|\frac{Z(\beta, t)}{Z(\beta)}\right|^2 = \frac{1}{Z(\beta)^2} \sum_{m,n} e^{-\beta(E_m + E_n)} e^{i(E_m - E_n)t} . \tag{2.3}$$

As in the case of the two-point function, the late time behavior of this quantity is generically complicated. One can make progress by taking the long-time average, where terms with oscillating phases average to zero and only terms with $E_m = E_n$

[2] The authors of [13] use this idea in a closely related context.

survive. It is given by

$$\lim_{T\to\infty}\frac{1}{T}\int_0^T dt\left|\frac{Z(\beta,t)}{Z(\beta)}\right|^2=\frac{1}{Z(\beta)^2}\sum_E N_E^2 e^{-2\beta E}\,, \tag{2.4}$$

where N_E is the degeneracy of the energy level E. If the spectrum has no degeneracies ($N_E=1$), the long-time average becomes

$$\lim_{T\to\infty}\frac{1}{T}\int_0^T dt\left|\frac{Z(\beta,t)}{Z(\beta)}\right|^2=\frac{Z(2\beta)}{Z(\beta)^2}\,. \tag{2.5}$$

Z generically scales as e^{aS} where S is the entropy and a is a positive constant. So (2.5) scales as e^{-aS}. In the holographic context S is the black hole entropy which scales as $1/g_s^2\sim 1/G_N$, where g_s and G_N are the string coupling and Newton constants of the bulk theory, so (2.5) is nonperturbative in the bulk coupling. For large black holes, S is given by the thermal entropy of the boundary field theory, and it scales with the number of degrees of freedom. In particular, we have $S\sim N^2$ in matrix theories like super-Yang-Mills (SYM) theory, and $S\sim N$ in vector theories like the Sachdev-Ye-Kitaev model [4,14]. Either way, the quantity (2.5) is non-perturbative in $1/N$.

Now, suppose we attempt to compute the left-hand side of (2.5) by making a coarse grained approximation. If we replace the discrete sum over states in (2.3) by an integral over a smooth density we find that the long-time average vanishes. In holography, by analytically continuing saddle points we also find disagreement with (2.5). (See Section 2.9 and also [15].) Therefore, by studying how the long-time decay of the partition function (or of the correlator) is avoided in gravity we are in fact probing the discreteness of the black hole spectrum — a basic

characteristic of its quantum nature.[3]

From the bulk perspective, Maldacena initially suggested that an instanton might be responsible for the analogous $O(e^{-aS})$ root-mean-square (RMS) height of the correlator. Barbon and Rabinovici [16] pointed out that such an instanton might not describe the details of the irregular long-time fluctuations expected in the correlator. Information loss in correlation functions was also studied in [17,18] in the context of $2d$ CFTs. These questions have been difficult to address in standard holographic contexts like $\mathcal{N}=4$ SYM, due to the difficulty in analyzing the chaotic boundary theory with sufficient precision.

The Sachdev-Ye-Kitaev (SYK) model [4, 14] is a good laboratory to explore these questions. It is a quantum mechanical model of N Majorana fermions with random q-fermion couplings that is soluble at large N. The theory is highly chaotic: at strong coupling it saturates [2, 19, 20] the chaos bound [5], a property that is characteristic of black holes in Einstein gravity [21–23]. It realizes a (highly curved) description of a "nearly AdS_2"/"nearly CFT_1" system [2, 3, 24–26]. As is the case for other vector models, there is an exact rewrite of the disorder-averaged model in terms of a functional integral over bilocal $O(N)$ singlet fields G, Σ that presumably are related to the bulk description.[4]

The SYK model has several other properties that make it useful in the study of late time properties. The average over the random couplings should rattle

[3] It is sometimes said that this problem is related to the question of why a black hole has finite entropy. Indeed, in standard QM, finite entropy implies a discrete spectrum, but we note that in disorder-averaged theories, or in a thermodynamic approximation, for example, one can effectively have a smooth but finite density of states.

[4] Higher dimensional versions of SYK have been constructed in [27, 28]. A supersymmetric generalization of the model has been constructed in [29]. A multiflavor version has been constructed in [30]. Other related work includes [31–34].

the energy eigenvalues sufficiently to make the rapidly oscillating terms in equations (2.1),(2.3) average to zero at a *fixed* time, making these quantities smooth functions of time. This makes them more amenable to study. In addition, the model is computationally simple enough that numerical methods can yield significant insight [35,36]. (After we had finished our numerical analysis the paper [37] appeared. It has significant overlap with our numerical results.)

One goal of this paper is to explore the late time behavior of the SYK model. We present numerous numerical results about such behavior in the model, and interpret them using a variety of analytic and conceptual arguments. One of our key findings is a close relationship between the late time behavior of the model and the behavior of random matrices.[5]

It is a widely held conjecture [40] that the spacing statistics of nearby energy levels in quantum chaotic systems should be well approximated by an appropriate random matrix ensemble. Since late times corresponds to small energy differences our result is a natural one.

Building on these observations we can make a plausible conjecture about the behavior of more complicated holographic systems, like the Type IIB AdS_5 / $\mathcal{N} = 4$ SYM system.

2.1.1 Summary of results

Here we give an outline of the paper and summarize the main results. In Section 2.2 we introduce the SYK model. Then in Section 2.3 we write down

[5] Another discussion of random matrices in black hole physics is [38]. Recent discussion of a connection between chaotic systems and random ensembles, including observables generalizing $\langle |Z(\beta, t)|^2 \rangle$, appears in [39].

the spectral form factor, which is given by $|Z(\beta,t)|^2/Z(\beta)^2$ averaged over the random couplings. At late times this quantity goes over to a plateau value given approximately by (2.5), which characterizes the discreteness of the spectrum. By numerically computing this quantity we find that its late time behavior exhibits an interesting feature, see Figure 2.1. Starting at $t = 0$, the spectral form factor first dips below its plateau value and then climbs back up in a linear fashion (we call this region the 'ramp'), joining onto the plateau. This behavior is readily explained if we approximate the SYK Hamiltonian by a Gaussian random matrix, as shown in Figure 2.2. Further evidence for the relation between the late time behavior and random matrix theory (RMT) is given in Section 2.3.1, where we show the relation between the choice of RMT ensemble (GUE, GOE, or GSE) and the detailed shape of the late time behavior in SYK. See Figure 2.4.

In Section 2.4 we make a digression to discuss the thermodynamic properties of SYK. We compute the entropy and energy numerically, and by extrapolating these results to infinite N we find excellent agreement with existing analytical calculations carried out in the large N limit. This serves as an incisive check both on our results and on existing analytic calculations.

In Section 2.5 we review the analytical origin of the ramp and plateau in RMT, and the relation of the ramp to the phenomenon of spectral rigidity. We show that the ramp can be understood as a perturbative effect in RMT (though not as a perturbative $1/N$ effect in SYK, as we explain).

In Section 2.6 we explain the early-time power-law decay in SYK visible in Figure 2.1. This is related to the low energy portion of the spectrum, dominant in the large N, large βJ limit, that is described by the Schwarzian theory of

reparametrizations. We argue this is exact in a double scaling limit. In the large N, large βJ limit a sector of the model [2, 3] is dual to a dilaton gravity [41, 42] black hole in AdS_2 . We argue that the subsequent linearly growing ramp and the plateau should survive in this limit, suggesting a connection between the late time behavior of black holes and random matrix theory.

In Section 2.7 we discuss a similar ramp that appears in SYK correlators. We work out the conditions under which the fermion two-point function exhibits the ramp/plateau structure of the spectral form factor, and check these results numerically.

In Section 2.8 we consider the behavior of the spectral form factor for a single realization of the random couplings. The motivation here is to make contact with theories such as Yang-Mills which do not involve an averaging over couplings. For a single realization the spectral form factor exhibits large fluctuations even at large N, but we argue that by time averaging (and no disorder averaging) the underlying ramp/plateau structure can be brought into view.

In Section 2.9 we make a connection with $\mathcal{N} = 4$ SYM, giving a preliminary estimate of the gravity saddle points that give the early-time decay of $|Z(\beta, t)|^2$. We also argue that there should be a subsequent long period of time where this quantity is growing and dominated by 'ramp' physics, folded against the coarse-grained density of states of the SYM theory.

We conclude and discuss future directions and ongoing work in Section 2.10.

Several appendices contain additional results and discussion.

In Appendix A.1 we review the particle-hole symmetry of the SYK model, whose properties depend on N mod 8 [36, 43].

In Appendix A.2 we discuss the double-scaled limit of SYK, where the disorder-averaged density of states can be computed exactly.

In Appendix A.3 we consider a toy model of the G, Σ path integral, which is an exact rewrite of the SYK model in terms of bosonic bilocal fields. We explain how the original fermionic behavior can arise from these bosonic variables.

In Appendix A.4 we again consider the G, Σ formulation of the model. We point out the existence of a family of subleading saddle points that show up both in the SYK model and in the integrable version $q = 2$ of it. We explain why this infinite family of saddle points does not significantly affect the thermodynamics of the model at large N.

In Appendix A.5 we further discuss these saddle points in the integrable $q = 2$ version of the SYK model, and show how they lead to a simple kind of random matrix theory behavior at late times.

In Appendix A.6 we make some preliminary remarks on the origin of the amplitude of the ramp in SYK.

In Appendix A.7 we present constraints on a simple single saddle point explanation of the ramp in SYK correlators.

Finally, in Appendix A.8 we present additional numerical data.

2.2 The Sachdev-Ye-Kitaev model

Consider N Majorana fermions ψ_a ($a = 1, \ldots, N$) in 0+1 dimensions that obey the algebra $\{\psi_a, \psi_b\} = \delta_{ab}$. The Hamiltonian is[6]

$$H = \frac{1}{4!} \sum_{a,b,c,d} J_{abcd} \psi_a \psi_b \psi_c \psi_d = \sum_{a<b<c<d} J_{abcd} \psi_a \psi_b \psi_c \psi_d \,. \tag{2.6}$$

The coupling tensor J_{abcd} is completely anti-symmetric, and each independent element is a random real number chosen from a Gaussian distribution with zero mean and variance given by $\sigma^2 = \frac{3!}{N^3} J^2$. The Hilbert space has dimension

$$L \equiv \text{dim. of Hilbert space} = 2^{N/2}, \tag{2.7}$$

and we set $J = 1$ for convenience.

In this work we mainly focus on the model with 4-fermion interactions, although we will sometimes discuss the generalization where the fermions interact in groups of q.

For N even it is often useful to implement the model using $N_d = \frac{N}{2}$ Dirac fermions c_i ($i = 1, \ldots, N_d$) by defining

$$\psi_{2i} = \frac{c_i + \bar{c}_i}{\sqrt{2}} \,, \quad \psi_{2i-1} = \frac{i(c_i - \bar{c}_i)}{\sqrt{2}} \,. \tag{2.8}$$

[6] We follow the conventions of [2] and specialize to $q = 4$, where q is the number of fermions interacting in each term of the Hamiltonian.

The Dirac fermions satisfy the algebra

$$\{c_i, \bar{c}_j\} = \delta_{ij}\,, \quad \{c_i, c_j\} = 0\,, \quad \{\bar{c}_i, \bar{c}_j\} = 0\,. \tag{2.9}$$

We can write down a fermion number charge given by $Q = \sum_{i=1}^{N_d} \bar{c}_i c_i$. The Hamiltonian (2.6) does not preserve this charge, but it does preserve charge parity (Q mod 2). Therefore, the Hamiltonian has two blocks corresponding to even and odd values of Q.

2.3 Spectral form factor

We define disorder-averaged analogs of the quantity in equation (2.3) as follows.

$$g(t;\beta) \equiv \frac{\langle Z(\beta,t) Z^*(\beta,t)\rangle_J}{\langle Z(\beta)\rangle_J^2}\,, \tag{2.10}$$

$$g_d(t;\beta) \equiv \frac{\langle Z(\beta,t)\rangle_J \cdot \langle Z^*(\beta,t)\rangle_J}{\langle Z(\beta)\rangle_J^2}\,, \tag{2.11}$$

$$g_c(t;\beta) \equiv g(t;\beta) - g_d(t;\beta)\,. \tag{2.12}$$

Here $\langle\cdot\rangle_J$ denotes the disorder average — the average over the ensemble of random couplings. $Z(\beta,t)$ was defined in (2.2). As discussed in the introduction, the late-time behavior of these quantities probes the discreteness of the spectrum, similar to the late-time behavior of two-point functions. Notice that we are working with annealed quantities, meaning that we are taking the disorder average separately in the numerator and denominator. This is in contrast with quenched quantities

such as $\langle |Z(\beta,t)|^2/Z(\beta)^2 \rangle_J$. The advantage of working with annealed quantities is that they require a finite number of replicas in analytic calculations (g requires two replicas, g_d requires just one), whereas quenched quantities require an arbitrary number of replicas.[7]

Now we present one of the central results of this work, $g(t)$ for the SYK model. In Figure 2.1 we present $g(t;\beta=5)$ for $N=34$, computed numerically.[8] Notice that $g(t)$ at early times does not simply join onto the late-time plateau, but instead dips below the plateau and then climbs back up. One goal of this work is to understand the source and implications of this behavior, and to estimate how prevalent it is both in SYK (for various values of the parameter βJ) and in quantum field theory in general.

Notice that $g(t)$ is smooth, and does not exhibit the large fluctuations that one expects at late times in a typical quantum theory. This is due to the disorder average, which smooths out the fluctuations exhibited by each realization of the random couplings. (Some fluctuations are apparent at late times, but these are an artifact due to the finite number of samples used in the computation. We will discuss this point further in Section 2.8.)

We will be discussing the curve $g(t)$ at length, so let us point out the main features in this plot and introduce some nomenclature. Starting with $t=0$, at early times the value of $g(t)$ drops quickly along what we will call the 'slope',

[7] Numerically, we find that the quenched and annealed versions of $g(t;\beta)$ remain well within a percent of each other for all times and values of β we considered, and the difference appears to decrease with N. (At infinite temperature the annealed and quenched quantities are in fact equal because $Z(\beta=0)=\mathrm{tr}(1)$ is independent of the random couplings.)

[8] All numerical results in this paper were computed by fully diagonalizing the SYK Hamiltonian for independently generated Gaussian random couplings, computing the relevant quantity, and then taking the mean.

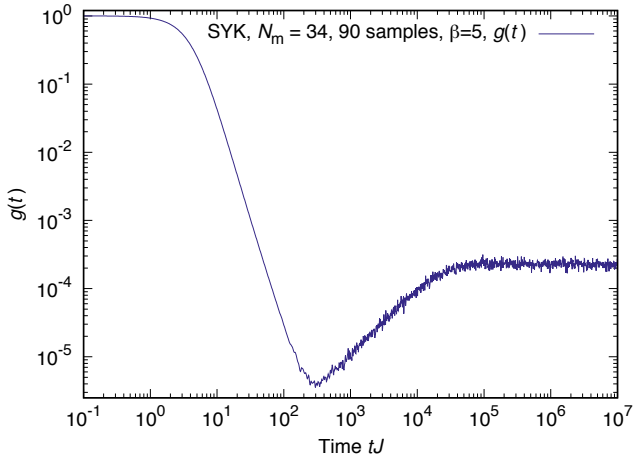


Figure 2.1: A log-log plot of SYK $g(t;\beta = 5)$, plotted against time for $N = 34$. Here we use the dimensionless combination tJ for time. Initially the value drops quickly, through a region we call the *slope*, to a minimum, which we call the *dip*. After that the value increases roughly linearly, $\sim t$, until it smoothly connects to a plateau around $tJ = 3 \times 10^4$. We call this increase the *ramp*, and the time at which the extrapolated linear fit of the ramp in the log-log plot crosses the fitted plateau level the *plateau time*. The data was taken using 90 independent samples, and the disorder average was taken for the numerator and denominator separately.

until it reaches a minimum at the 'dip time' t_d. Next comes a period of linear growth that we will call the 'ramp'. It ends at the plateau time t_p, and beyond this we have an almost constant value of $g(t)$ that we call the 'plateau'. The plateau height is equal to the long-time average of $g(t)$. On the plateau only the $E_n = E_m$ terms in the sum (2.3) survive, and the height of the plateau is $2Z(2\beta)/Z^2(\beta) \sim e^{-aS}$, in accordance with (2.4). The factor of 2 is due to a 2-fold degeneracy in the spectrum (see Appendix A.1).

Quantities such as g, g_d, and g_c are studied extensively in the field of quantum chaos. In particular, $g(t)$ (typically used with $\beta = 0$) is called the *spectral form factor* and it is a standard diagnostic of the pair correlation function of energy

eigenvalues. We will often refer to $g(t)$ by this name. It supplies information about the correlations of eigenvalues at different energy separations.[9]

One of the basic conjectures in the field of quantum chaos is that the fine grained energy eigenvalue structure of a chaotic system is the same as that of a random matrix chosen from one of the standard Dyson ensembles [44]: Gaussian Unitary Ensemble (GUE), Gaussian Orthogonal Ensemble (GOE), or Gaussian Symplectic Ensemble (GSE). (For reviews, see [40, 45].) The particular ensemble to use depends on the symmetries of the original Hamiltonian. Random matrix theory can then be used to compute certain quantities (such as the spectral form factor) that are sensitive to eigenvalue correlations. You, Ludwig and Xu [36] first discussed the quantum chaotic properties of SYK by studying the distribution of spacings between nearest-neighbor energy levels, another standard quantum chaos observable. They showed that the distribution is consistent with RMT predictions.

In Figure 2.2 we present $g(t; \beta = 5)$ for the GUE ensemble of matrices of rank $L_{\text{RMT}} = 2^{12}$, computed numerically, with a normalization such that the eigenvalues typically lie in the range $-2 < \lambda < 2$ (see (2.24)). At $\beta = 0$ the height of the plateau is of order $1/L_{\text{RMT}}$ and the plateau time is at t of order L_{RMT}, the inverse mean level spacing.

Note the similarity between the RMT result and the SYK result, and in particular the presence of the ramp and the plateau. We will argue that the late-time behavior of the spectral form factor in SYK can be explained by random matrix

[9] The spectral form factor contains information about the pair correlation between well-separated eigenvalues that the (perhaps more familiar) diagnostic of the nearest-neighbor energy spacing distribution does not. Conversely, the nearest-neighbor level spacing distribution contains information about multi-point correlation functions of nearby eigenvalues that the spectral form factor does not.

theory. The early time behavior of RMT differs from SYK, although it is not obvious from the plots. The typical eigenvalue density has different dependence on energy in the two systems, which leads to somewhat different initial decays. Moreover, at early times RMT is governed by a perturbative expansion in $1/L$, while SYK is governed by an expansion in $1/N$. On the other hand, at times well beyond the dip, $g(t)$ is determined by eigenvalue correlations on scales much smaller than the total width of the spectrum, and there one expects to find agreement between SYK and RMT.

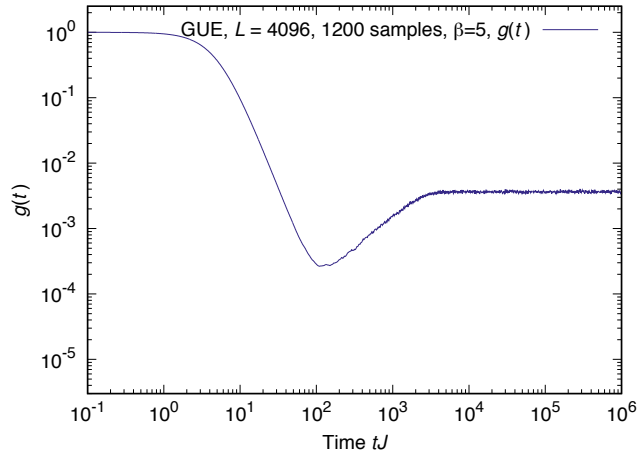


Figure 2.2: A log-log plot of $g(t;\beta = 5)$ against time for GUE random matrices, dimension $L = 2^{12}$. A dip, ramp and plateau structure similar to Fig. 2.1 is apparent.

What is the physical origin of the ramp in RMT? Eigenvalues of generic matrices repel, so near degeneracies are extremely unlikely. This causes the plateau. The time of onset of the plateau is determined by the scale of near neighbor eigenvalue spacings. The ramp, though, is due to the repulsion between eigenvalues that are far apart in the spectrum. This repulsion, when balanced against the effects that keep the energy finite, gives rise to a very rigid eigenvalue structure.

This phenomenon is referred to as long-range spectral rigidity [40, 44, 45]. More quantitatively, if δE_n denotes the deviation of an energy from its average value, then at leading order $\langle \delta E_n \delta E_m \rangle \sim \log |n-m|$. For comparison, if the eigenvalues formed a one dimensional crystal with harmonic near neighbor interactions, then $\langle \delta E_n \delta E_m \rangle \sim |n-m|$, a much less rigid behavior [40, 44, 45]. The $\log |n-m|$ form, after suitable processing we will discuss below, accounts for the linear behavior of the ramp. The ramp lies below the plateau because repulsion causes the eigenvalues to be anticorrelated.

2.3.1 The ramp and the eightfold way

We now present further evidence of the relation between random matrix theory and the presence of the ramp in the SYK spectral form factor.

The Hamiltonian of a chaotic theory is generally believed to resemble a random matrix when studied at sufficiently fine energy resolution. One basic property of random matrices is their nearest-neighbor level statistics, namely the distribution of the distance s between pairs of neighboring energy levels [46].[10] The nearest-neighbor statistics of an integrable theory follow an exponential distribution e^{-s}, while those of a chaotic theory generally follow one of the three reference ensembles GUE, GOE, and GSE. The particular ensemble depends on the symmetries of the Hamiltonian.

You, Ludwig and Xu [36] studied the nearest-neighbor level spacing distribution in SYK. They made the important point that all three Gaussian RMT

[10] More precisely, one considers the distribution of spacings between *unfolded* energy levels [47]. These are the levels one obtains by making a change of variables such that the mean level spacing becomes one everywhere. For further details, see [45].

ensembles are implemented in the model as we now review.

The SYK model has a particle-hole symmetry given by [35, 36, 43]

$$P = K \prod_{i=1}^{N_d} (\bar{c}_i + c_i) \,, \tag{2.13}$$

where K is an anti-linear operator. The properties of this operator determine the class of RMT statistics of each charge parity sector of the Hamiltonian. In particular, the statistics are determined by the value of (N mod 8) as follows (see Appendix A.1 for details).

- When N mod $8 = 2$ or 6, the symmetry P maps the even and odd parity sectors to each other. Individual sectors do not have any anti-linear symmetry, and the corresponding ensemble of each sector is GUE.

- When N mod $8 = 0$, P maps each sector to itself and $P^2 = 1$. The corresponding ensemble is GOE.

- When N mod $8 = 4$, P again maps each sector to itself but now $P^2 = -1$. The corresponding ensemble is GSE.

Figure 2.3 shows the nearest-neighbor statistics of SYK with $N = 30, 32$, and we see excellent agreement with RMT predictions.

While the nearest-neighbor spacing distribution is sensitive to correlations between adjacent energy levels, the spectral form factor probes correlations between energy levels at larger separations. The t parameter in $g(t)$ determines the scale of the energy differences being probed. As discussed above, beyond the plateau time only individual energy levels are probed, while at earlier times (and in particular

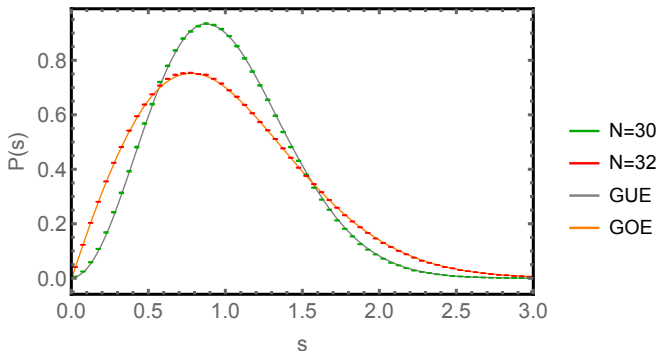


Figure 2.3: Unfolded nearest-neighbor level spacing distribution for SYK vs. RMT. Here s is measured in units of the mean spacing. Semi-analytical exact large L results (correcting the Wigner surmise) for the RMT $P(s)$ are available [48, 49], but we computed the RMT curves from $L = 12870$ exact diagonalization data.

on the ramp) $g(t)$ is sensitive to correlations between levels that are much farther apart than the mean level spacing. The structure of these correlations depends on the ensemble. The three RMT ensembles all exhibit a ramp and a plateau but with slightly different shapes: In GUE the (unfolded) ramp and plateau connect at a sharp corner, in GOE they connect smoothly, and in GSE they connect at a kink.[11]

Figure 2.4 shows $g(t)$ at $\beta = 0, 1, 5$ for various values of N. The corresponding RMT ensembles are

N	16	18	20	22	24	26	28	30	32	34
class	GOE	GUE	GSE	GUE	GOE	GUE	GSE	GUE	GOE	GUE

The shape of the ramp in each case agrees with the RMT prediction outlined above. In particular, the kinks visible for $N = 20, 28$ are a signature feature of the ramp in the GSE ensemble. For $N = 34$ (GUE) a careful comparison that confirms the RMT ramp shape is described in Section 2.6. As an initial test we

[11] See, for instance, Figure 10 in [45].

fitted the ramp at times well before the plateau time (where unfolding effects discussed in Section 2.6 become significant). We found a power behavior agreeing with the expected GUE behavior $g(t) \sim t^1$ to within a few percent. These are strong pieces of evidence that the ramp structure in SYK can be attributed to random matrix theory.

For $\beta = 0$ the early time behavior exhibits oscillations, which will not play a role in this work. The oscillations are due to the fact that, at infinite temperature, the spectral form factor is sensitive to the hard edges at both ends of the energy spectrum.

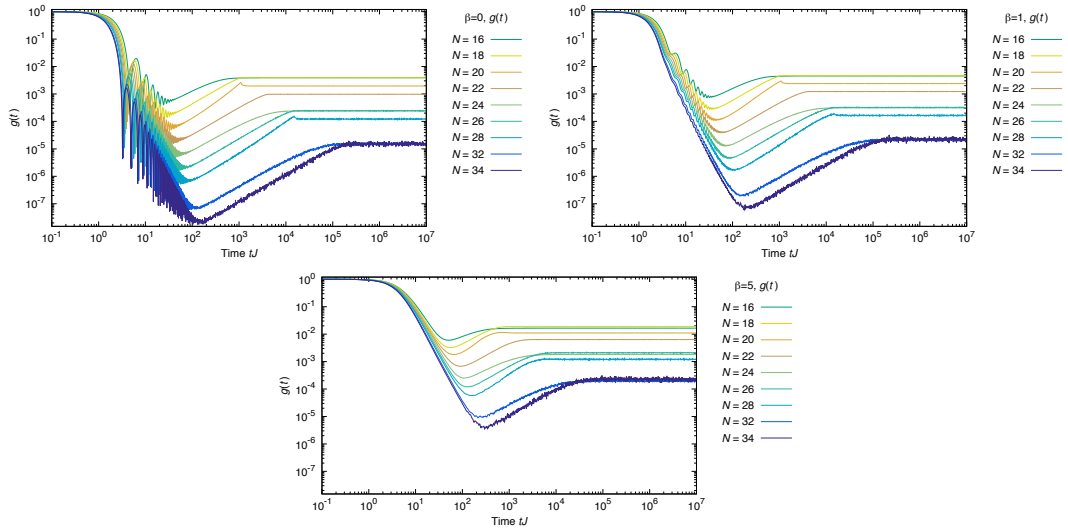


Figure 2.4: SYK $g(t, \beta)$ with $\beta = 0, 1, 5$ and various N values. The value at late times, which is equal to plateau height g_p, matches with $N_E Z(2\beta)/Z(\beta)^2$ as discussed in Appendix A.8.2. Here N_E is the eigenvalue degeneracy, 2 for (N mod 8) $\neq 0$ and 1 for (N mod 8) $= 0$. As explained in the main text, the shape of the ramp and the plateau depends on the symmetry class, and the agreements with the counterparts in the RMT with GUE, GOE, and GSE are good. The numbers of samples are 1 200 000 ($N = 16$), 600 000 ($N = 18$), 240 000 ($N = 20$), 120 000 ($N = 22$), 48 000 ($N = 24$), 10 000 ($N = 26$), 3 000 ($N = 28$), 914 ($N = 30$), 516 ($N = 32$), 90 ($N = 34$).

Let us now consider the plateau heights of Figure 2.4 in detail. They are equal to the time-average value of $g(t)$, which at $\beta = 0$ is given by (2.4)

$$\frac{\sum_E N_E^2}{L^2}\,. \tag{2.14}$$

Here N_E is the degeneracy of energy level E. As explained in Appendix A.1, the SYK spectrum has a double degeneracy ($N_E = 2$) when (N mod 8) $\neq 0$ due to the particle-hole symmetry, leading to a plateau height of $2/L$ at $\beta = 0$. When (N mod 8) $= 0$ there is no protected degeneracy, and in those cases the plateau height is $1/L$. These facts are consistent with the pattern of plateau heights exhibited by Figure 2.4. In particular, they explain why the plateaus of $N = 16, 24, 32$ are reduced by a factor of 2 compared with the rest.

One important consequence of Figure 2.4 is that it allows us to learn about the large N behavior of the ramp. As we go to larger N the dip time grows quickly, but the plateau time grows even faster, resulting in a more and more prominent ramp. (For further discussion of the numerical evidence, see Appendix A.8.2.) We are led to the reasonable conjecture that the ramp is a feature of the large N theory, and that the dip time is a new time scale in the theory. In Section 2.6 we will present an analytic argument that supports this conclusion.

2.4 Thermodynamics of the SYK model

In this section we compute the thermodynamic properties of SYK numerically, and extrapolate to the large N limit. We find excellent agreement with existing analytic results, both for the infinite N limit and for the leading $O(1/N)$ correc-

tion. This serves as an important cross-check both on our results and on existing results.

We begin with a brief review of the known analytic results. There is an exact rewrite of the SYK model in terms of bi-local anti-symmetric variables $G(\tau_1, \tau_2)$ and $\Sigma(\tau_1, \tau_2)$. The path integral is given in Euclidean time by [14, 50]

$$Z = \int DGD\Sigma\, e^{-I}\,, \tag{2.15}$$

$$\frac{I}{N} = -\frac{1}{2}\log\det(\partial_\tau - \Sigma) + \frac{1}{2}\int_0^\beta d\tau_1 d\tau_2 \left[\Sigma(\tau_1,\tau_2)G(\tau_1,\tau_2) - \frac{J^2}{q}G^q(\tau_1,\tau_2)\right]. \tag{2.16}$$

We remind the reader that we set $q = 4$ in most of our analysis. The action (2.16) is obtained by performing the disorder average over couplings J_{ijkl}, introducing a Hubbard-Stratonovich field for the fermion bi-linear, and integrating out the fermions [14, 50]. In particular, $G(\tau_1, \tau_2)$ should be thought of as the fermion bi-linear $\frac{1}{N}\sum_{a=1}^N \psi_a(\tau_1)\psi_a(\tau_2)$ and $\Sigma(\tau_1, \tau_2)$ as a Lagrange multiplier enforcing this identification. To compute $\langle Z(\beta + it)Z(\beta - it)\rangle_J$ we need two copies (called replicas) of the fermion fields labelled by replica indices $\alpha, \beta = 1, 2$. G, Σ become $G_{\alpha\beta}, \Sigma_{\alpha\beta}$. The convergence of (2.15) is manifest with a contour choice described in Appendix A.3.

To solve the theory at large N one now writes the saddle point equations for the bi-local fields.

$$\frac{1}{G(\omega)} = -i\omega - \Sigma(\omega)\,, \quad \Sigma(\tau) = J^2 G^{q-1}(\tau)\,. \tag{2.17}$$

The first equation is in frequency space, and the second is in Euclidean time. These equations can be solved analytically in the limits $\beta J \to 0$ and $\beta J \to \infty$ [51], and can be solved numerically for arbitrary values of βJ. Plugging the result back in (2.16) gives the large N thermal free energy [52]. Certain perturbative $1/N$ corrections to the free energy have also been computed [2, 53].

At finite N we compute the mean energy and other thermodynamic quantities numerically by fully diagonalizing the Hamiltonian. To make contact with the analytic calculation, we extrapolate the numerical results to large N as follows. At fixed temperature T we compute $\langle E(T)\rangle/N$ at different N values and fit to a polynomial in $1/N$ of degree 2. The leading $O(N^0)$ coefficient is then the infinite N result, the next term is the $1/N$ correction, and so on.

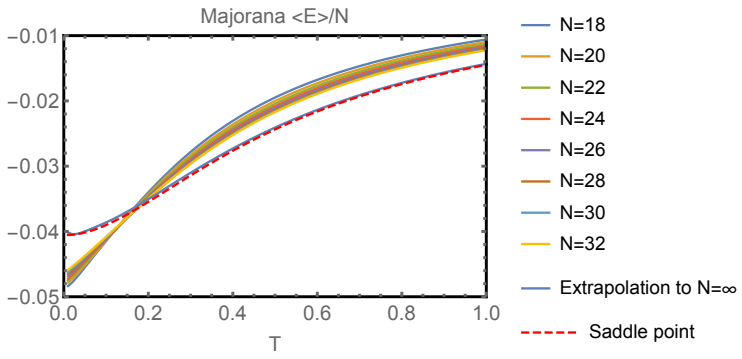


Figure 2.5: Shown are SYK thermodynamic $\langle E(T)\rangle/N$ for different values of N, computed by exact diagonalization. We also plot the point-wise extrapolation obtained by fitting the eight values of N to a three-tem expansion in $1/N$ and taking the leading term. This is almost indistinguishable from the exact large N result obtained by solving the Schwinger-Dyson equations numerically.

Figure 2.5 shows the mean energy extrapolated to infinite N, compared with the result obtained from a direct solution of the large N saddle point equations. We find excellent agreement between the two methods of computation, although even at $N = 32$ (the largest value considered here) the result is not close to the

infinite N answer. The mean energy can be written at low temperature as

$$\langle E(T)\rangle = N\epsilon_0 + aT + \frac{Nc}{2}T^2 + Nc_2T^3 + \cdots \tag{2.18}$$

The normalization is such that all coefficients scale as $O(N^0)$. The coefficients c (the large N specific heat) and a have been computed in the large N theory[12] and are given by

$$\epsilon_0 \approx -0.0406\,, \quad a = \frac{3}{2}\,, \quad \frac{c}{2} \approx 0.198\,. \tag{2.19}$$

The coefficient c_2 has not been reported in the literature, but we believe it should be $c_2 = -0.419$.[13] Notice that the linear term in (2.18) is subleading in $1/N$. This must be the case because this term corresponds to a $\log(T)$ term in the entropy, which becomes negative at finite temperature. Let us now compare these coefficients to the extrapolated numerical results:[14]

$$\frac{E}{N} = -0.04 - 0.0025T + 0.22T^2 - 0.52T^3 + 0.37T^{3.77}\,. \tag{2.21}$$

We see that a is suppressed at infinite N as expected, while c is within fifteen

[12] The coefficient a was computed in [2,53] from a one-loop fluctuation correction to the large N saddle, or equivalently from summing diagrams formed by bending ladder diagrams around into a loop.

[13] This is based on a conjectured $1/\beta^2$ term in the free energy, which in the notation of [2] reads

$$\frac{\log Z}{N} = \#\beta\mathcal{J} + s_0 + \frac{2\pi^2\alpha_S}{\beta\mathcal{J}} - \frac{2\pi^2\alpha_S\alpha_K}{(\beta\mathcal{J})^2|k_c'(2)|} + \ldots \tag{2.20}$$

[14] We included a $T^{3.77}$ term to account for the first nontrivial operator dimension in the model [2]. Surprisingly, the fit agrees with large N results slightly better if we replace this with a T^4 term.

percent of the expected value (2.19). Next, we fit the $1/N$ piece of the extrapolated energy and find

$$-0.23 + 1.6T - 3.4T^2 + 2.9T^3 \,. \tag{2.22}$$

Here the fitted value of $a = 1.6$ is fairly close to the expected value $a = \frac{3}{2}$.

Next, Figure 2.6 shows the entropy extrapolated to infinite N. We again find excellent agreement with a direct infinite N calculation. At low temperature the entropy is given by

$$S(T) = N s_0 + a \log(T) + NcT + \cdots \,. \tag{2.23}$$

Here $s_0 \approx 0.2324 \approx \frac{1}{2}\log(1.592)$ is the analytic zero-temperature entropy density (in the large N limit). Notice that the numerical extrapolation correctly captures the large N zero-temperature entropy, even though at any fixed N the entropy goes to zero as $T \to 0$.

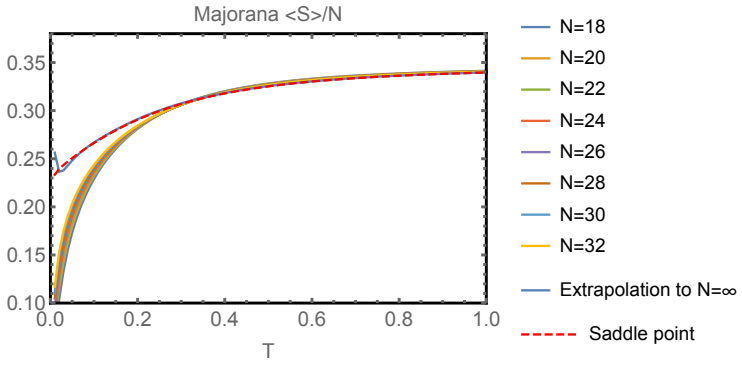


Figure 2.6: SYK thermodynamic $S(T)/N$, analyzed in the same way as Fig. 2.5.

2.5 Spectral form factor in random matrix theory

In this section we review properties of the spectral form factor in the GUE random matrix ensemble [40,45]. We derive two of the main features of Figure 2.2: the late time behavior of the slope and the early time behavior of the ramp. Both are described by power laws, and from there we get an estimate of the dip time in RMT.

Consider the GUE ensemble of Hermitian matrices M of rank L, with ensemble averaging defined by

$$\mathcal{Z}_{\text{GUE}} = \int \prod_{i,j} dM_{ij} \exp\left\{ \left(-\frac{L}{2} \operatorname{tr}(M^2) \right) \right\}. \tag{2.24}$$

In this context, the matrix M is analogous to the SYK Hamiltonian, and the rank L corresponds to the dimension of the Hilbert space. One important difference is that the natural perturbative parameter in SYK is $1/N$, whereas in RMT we typically expand in $1/L \sim e^{-N}$.

The partition function for a given realization of M is defined by

$$Z(\beta, t) \equiv \operatorname{tr}\left(e^{-\beta M - iMt}\right) . \tag{2.25}$$

The spectral form factor g and the related quantities g_d and g_c are then defined by (2.10)-(2.12), where the average $\langle \cdot \rangle_J$ over the random couplings is replaced by

the average $\langle\cdot\rangle_{\text{GUE}}$ over random matrix elements, given by (2.24).[15]

Let us diagonalize M and change variables from its matrix elements to its eigenvalues and a unitary change of basis. This introduces a Jacobian that describes repulsion between eigenvalues. In the large L limit the eigenvalues can be described by a density ρ. We will use ρ for the physical density, and $\tilde{\rho}$ for the unit normalized density:

$$\int d\lambda \rho(\lambda) = L, \qquad \int d\lambda \tilde{\rho}(\lambda) = 1, \qquad \rho(\lambda) = L\tilde{\rho}(\lambda). \tag{2.27}$$

Replacing the individual eigenvalues λ_i by $\tilde{\rho}(\lambda)$, one obtains[16]

$$\mathcal{Z}_{\text{GUE}} = \int D\tilde{\rho}(\lambda)\, e^{-S}\,, \quad S = -\frac{L^2}{2}\left\langle \lambda^2 \right\rangle + L^2 \left\langle \log|\lambda_1 - \lambda_2| \right\rangle\,. \tag{2.28}$$

The large L saddle point of the above is given by the Wigner semicircle law,

$$\langle \tilde{\rho}(\lambda)\rangle_{\text{GUE}} = \tilde{\rho}_s(\lambda) \equiv \frac{1}{2\pi}\sqrt{4-\lambda^2}\,. \tag{2.29}$$

The physical eigenvalue density is given by $\langle\rho(\lambda)\rangle = L\tilde{\rho}_s(\lambda)$. Notice that the average eigenvalue spacing is of order $1/L$.

We now turn to discuss the slope and ramp that appear in the spectral form factor, shown in Figure 2.2. Roughly speaking, $g(t)$ is dominated by the discon-

[15] For example, for the partition function we have

$$\langle Z(\beta,t)\rangle_{\text{GUE}} = \frac{1}{\mathcal{Z}_{\text{GUE}}}\int dM_{ij}\, e^{-\frac{L}{2}\,\text{tr}(M^2)}\,\text{tr}\left(e^{-\beta M - iMt}\right)\,. \tag{2.26}$$

[16] The normalization of $\rho(\lambda)$ should be imposed (for example) by a Lagrange multiplier. The resulting saddle point equations are solved subject to this constraint.

nected piece $g_d(t)$ before the dip time, and by the connected piece $g_c(t)$ after the dip time. We will discuss each in turn.

The leading large L behavior of $Z(\beta, t)$ follows from the semicircle law. Working for simplicity at infinite temperature, we have

$$\langle Z(\beta = 0, t)\rangle_{\text{GUE}} = \int_{-2}^{2} d\lambda L \tilde{\rho}_s(\lambda) e^{-i\lambda t} = \frac{L J_1(2t)}{t} \,. \tag{2.30}$$

Here J_1 is a Bessel function of the first kind. At late times we find that the partition function decays as $L/t^{3/2}$, and therefore at late times we have

$$g_d(t) \equiv \frac{|\langle Z(0,t)\rangle_J|^2}{L^2} \sim \frac{1}{t^3} \,. \tag{2.31}$$

This is true also at finite temperature. Before the dip time, the spectral form factor $g(t)$ is dominated by the disconnected part $g_d(t)$. Therefore, the late time decay of $g(t)$ before the dip time is also proportional to $1/t^3$. This particular power is a consequence of the fact that the mean eigenvalue density (2.29) vanishes as a square root near the edge of the spectrum.

2.5.1 The ramp and the dip time

We now review how to derive the presence of a ramp in RMT. We focus for simplicity on the connected spectral form factor $g_c(t; \beta = 0)$, and show that $g_c(0, t) \sim \frac{t}{L^2}$ at times $1 \ll t \leq L$. Beyond the dip time, $g(t)$ and $g_c(t)$ are almost equal, both exhibiting the ramp/plateau structure. However, for g_c the ramp

extends to very early times, giving better perturbative control.[17]

The connected spectral form factor can be written as

$$g_c(t;0) = \int d\lambda_1 d\lambda_2 \, R_2(\lambda_1,\lambda_2) e^{i(\lambda_1-\lambda_2)t} \,, \tag{2.32}$$

$$R_2(\lambda_1,\lambda_2) \equiv \langle \delta\tilde{\rho}(\lambda_1)\delta\tilde{\rho}(\lambda_2)\rangle_{\rm GUE} \,. \tag{2.33}$$

Here R_2 is the connected pair correlation function of the unit-normalized density $\tilde{\rho}$, and $\delta\tilde{\rho}(\lambda) \equiv \tilde{\rho}(\lambda) - \tilde{\rho}_s(\lambda)$ is the fluctuation around the mean eigenvalue density $\tilde{\rho}_s(\lambda)$ given by the semicircle law (2.29). A basic result of RMT is that, near the center of the semicircle, $R_2(\lambda_1,\lambda_2)$ is given by the square of the *sine kernel* [48, 49, 54] plus a delta function at coincident points:

$$R_2(\lambda_1,\lambda_2) = -\frac{\sin^2\left[L(\lambda_1-\lambda_2)\right]}{\left[\pi L(\lambda_1-\lambda_2)\right]^2} + \frac{1}{L\pi}\delta(\lambda_1-\lambda_2). \tag{2.34}$$

Fourier transforming as in (2.32) gives

$$g_c(t) \sim \begin{cases} t/(2\pi L^2)\,, & t < 2L \\ 1/(\pi L)\,, & t \geq 2L \end{cases} . \tag{2.35}$$

This explains the observed behavior in Figure 2.2 beyond the dip time: There is a ramp up to the plateau time $2L$, and a constant plateau value beyond.[18] The ramp lies below the plateau because the eigenvalues are anticorrelated as reflected

[17] In fact, $1/L$ perturbation theory remains valid up to times $t \sim \epsilon L$ where ϵ is a small L-independent parameter.

[18] Our analysis here only applies to the contribution from eigenvalues near the center of the semicircle, where the mean density is L/π. We will show how to include regions with different mean densities in (2.42). Brezin and Hikami derive remarkable nonperturbative formulas for $g(t)$ in [55, 56].

in the minus sign in (2.34).

This is a good explanation of the ramp and plateau, but it requires an appeal to the sine kernel. In fact, one can derive the ramp in a more basic way without knowing about the sine kernel. Notice that the initial linear time dependence of the ramp can be obtained by approximating the sine kernel by

$$R_2(\lambda_1, \lambda_2) \approx -\frac{1}{2(\pi L(\lambda_1 - \lambda_2))^2} \,. \tag{2.36}$$

We now review how to derive this perturbatively from the action (2.28) following Altshuler and Shklovskii [57]. Writing $\tilde{\rho} = \tilde{\rho}_s + \delta\tilde{\rho}$ and expanding the action (2.28) about the saddle point, we find the quadratic term

$$\delta S = -L^2 \int d\lambda_1 \, d\lambda_2 \, \delta\tilde{\rho}(\lambda_1)\delta\tilde{\rho}(\lambda_2) \log|\lambda_1 - \lambda_2| \,. \tag{2.37}$$

We can now carry out the Gaussian integral to determine the two-point function (2.33). We go to Fourier space $\delta\tilde{\rho}(\lambda) = \int \frac{ds}{2\pi} \, \delta\tilde{\rho}(s) \exp(is\lambda)$ and find

$$\delta S = \frac{L^2}{2} \int ds \, \delta\tilde{\rho}(s) \frac{1}{|s|} \delta\tilde{\rho}(-s) \,. \tag{2.38}$$

Notice that long-wavelength fluctuations of ρ are strongly suppressed: This is the

spectral rigidity referred to in RMT.[19] Then we find

$$\begin{aligned}\langle \delta\tilde{\rho}(\lambda_1)\delta\tilde{\rho}(\lambda_2)\rangle &= \frac{1}{4\pi^2 L^2}\int ds e^{i(\lambda_1-\lambda_2)s}|s| + O(L^{-4}) \\ &= -\frac{1}{2(\pi L(\lambda_1-\lambda_2))^2} + O(L^{-4})\,. \end{aligned} \tag{2.39}$$

A calculationally more efficient way of studying $g(t)$ in RMT is the formalism developed by Brezin and Zee [58] which uses standard 't Hooft large L perturbation theory to compute the double resolvent of M. We discuss this technology in Appendix A.6.

Equating the slope (2.31) and the ramp (2.35) gives the RMT dip time $t_d \sim \sqrt{L}$, exponential in the "entropy" $\log L$. We find that the ratio $t_p/t_d \sim \sqrt{L}$, also exponential in the entropy, and therefore the ramp in the RMT spectral form factor survives in the large L limit.

This derivation makes it clear that (2.39) is a perturbative result in RMT at order $1/L^2$. Its contribution to $g_c(t)$ is proportional to t/L^2, capturing the ramp part of (2.35). In other words the ramp is a perturbative RMT effect. By contrast, the plateau is not.[20] Indeed, the appearance of the plateau depends on the oscillating factor in the more exact sine kernel (2.34) expression, which can be obtained from a RMT instanton expression $e^{-2LE_{\text{imag}}}$ with imaginary energy [59,60]. The oscillating term comes from continuing to real energy and extracting the appropriate part of the result.

[19] By observing that the local eigenvalue density is the inverse of the level spacing one can read off from the following result the $\langle \delta E_n \delta E_m\rangle \sim \log|E_n - E_m|$ signature of spectral rigidity discussed earlier.

[20] In GUE the ramp is the full perturbative result, while in other RMT ensembles (such as GOE and GSE) the ramp receives higher-order perturbative corrections. Non-perturbative corrections to the ramp exist in all cases.

This has important consequences for the application to the SYK model. For SYK, $L = 2^{N/2}$ so $1/L \sim e^{-aN}$. Therefore, perturbative RMT effects are nonperturbative in SYK, of order e^{-2aN}. Nonperturbative RMT effects of order e^{-L} are of order $\exp\left(-e^{aN}\right)$, an extremely small nonperturbative effect.

2.6 Spectral form factor in the SYK model

The presence of the ramp in the results of Section 2.3 suggests that the SYK model possesses spectal rigidity, even for eigenvalue spacings far larger than the mean nearest-neighbor spacing. By combining this assumption with coarse-grained features of the large N spectrum, we reproduce reasonably well the $g(t)$ curve obtained from exact diagonalization.

First, let us explain how an assumption of spectral rigidity produces the ramp observed in $g(t)$. Starting with the general definition of $\langle ZZ^* \rangle$,

$$\langle Z(\beta + it)Z(\beta - it)\rangle = \int d\lambda_1 d\lambda_2 \langle \rho(\lambda_1)\rho(\lambda_2)\rangle e^{-\beta(\lambda_1+\lambda_2)} e^{-i(\lambda_1-\lambda_2)t}, \tag{2.40}$$

it is convenient to define $x = \lambda_1 - \lambda_2$ and $E = \frac{\lambda_1+\lambda_2}{2}$. Notice that in this expression and below, we are using ρ, the physical eigenvalue density, normalized so $\int d\lambda\, \rho = L$.

Now, for late times we assume that the integral is dominated by regions where x is sufficiently small that we can approximate the density-density correlator by

GOE, GSE, or GUE statistics. For simplicity, we take GUE statistics

$$\langle \rho(\lambda_1)\rho(\lambda_2)\rangle = \langle\rho(E)\rangle\delta(x) + \langle\rho(\lambda_1)\rangle\langle\rho(\lambda_2)\rangle\left(1 - \frac{\sin^2\left[\pi\langle\rho(E)\rangle x\right]}{\left[\pi\langle\rho(E)\rangle x\right]^2}\right), \tag{2.41}$$

which leads to[21]

$$\langle Z(\beta+it)Z(\beta-it)\rangle = |\langle Z(\beta+it)\rangle|^2 + \int dE\, e^{-2\beta E}\min\left\{\frac{t}{2\pi}, \langle\rho(E)\rangle\right\}. \tag{2.42}$$

Eq. (2.42) can be interpreted as follows: we approximate the spectrum by bands over which $\rho(E)$ varies very little. From each band, we get a GUE ramp. The integral over energy in (2.42) is simply summing up these individual ramps which then yields another smoothed ramp. This is the inverse of an "unfolding" process. In a theory with many degrees of freedom, we expect the integral over E to be strongly peaked around a maximum. In general, the location of this maximum will depend on time. The ramp will join the plateau at the time $t_p = e^{S(2\beta)}$, where the energy that maximizes the integral is simply $E(2\beta)$, the energy that dominates the canonical ensemble at inverse temperature 2β. One can check that the derivative of $g_c(t)$ will smoothly approach zero at t_p, giving a C^1 transition onto the plateau even though individual bands have a kink.

One would like to apply (2.42) to SYK, but there is an important subtlety. The second term in (2.42) should be understood as exponentially smaller than the first, as long as t is not too large. In SYK, we also expect correlations be-

[21] We should make a few comments about this formula. First, if the local statistics are GOE or GSE, then we would replace the ramp function in (2.42) by the appropriate spectral form factor. Second, in cases where the spectum is uniformly d-fold degenerate, we should multiply the ramp term by d^2 and divide $\langle\rho(E)\rangle$ inside the second term by d.

tween eigenvalues that are only power-law suppressed by N (more precisely of order $1/N^q$). One source of such fluctuations would be the overall scale of the Hamiltonian, which varies from J configuration to J configuration. Such terms would dominate over the ramp contribution at short times. However, we might hope that these $1/N^q$ terms will always be smaller than either the first term or the second term in (2.42), so the formula still gives a reasonable picture of SYK. We will return to this point below.

Let us now attempt to evaluate (2.42) for large N SYK. First we discuss the disconnected first term. We can numerically evaluate the large N saddle point that determines $\langle Z(\beta+it)\rangle$, but for large values of $\beta+it$, we also need to consider fluctuations about this saddle. There are a set of modes that become soft for large $\beta+it$, which can be captured by the partition function of the effective Schwarzian derivative theory [2, 22]:

$$Z_{Sch}(\beta) = \int \frac{D[\tau(u)]}{SL(2,R)} \exp\left[-\frac{\pi N \alpha_S}{\beta \mathcal{J}} \int_0^{2\pi} du \left(\frac{\tau''^2}{\tau'^2} - \tau'^2\right)\right]. \tag{2.43}$$

Here, $0 < u < 2\pi$ is the physical time variable of the model, and $\tau(u)$ is a reparametrization of the thermal circle. The parameter $\mathcal{J}$ sets the scale of the Hamiltonian in a way appropriate for general values of q, and α_S is a numerical coefficient that depends on q; these are related to the specific heat c by $c = \frac{4\pi^2\alpha_S}{\mathcal{J}}$. The classical and one-loop contributions to this action have been studied previously [2], with the result

$$Z_{Sch}^{1-loop}(\beta) = \frac{\#}{(\beta\mathcal{J})^{3/2}} \exp\left(\frac{2\pi^2 N \alpha_S}{\beta \mathcal{J}}\right). \tag{2.44}$$

However, notice that when we continue to large values of $\beta + it$, the coefficient multiplying the action (2.43) becomes small, and $\tau(u)$ will have large fluctuations. Naively, this invalidates a perturbative analysis, making it difficult to evaluate Z. In fact, with the correct measure, the theory turns out to be one-loop exact. We will present a somewhat indirect derivation of this fact. A direct proof is also possible [61].

Our derivation is based on an intermediate step where we think about the SYK model for large values of q. Then the coefficient in the action becomes [2]

$$\frac{\pi N \alpha_S}{\beta \mathcal{J}} \to \frac{\pi}{4\beta\mathcal{J}} \cdot \frac{N}{q^2}. \tag{2.45}$$

In particular, the coefficient is only a function of $\frac{N}{q^2}$. We can therefore take a "double-scaled" limit of large N and large q with $\frac{N}{q^2}$ held fixed. It is clear that the Schwarzian part of the theory will survive in this limit, but the rest of the SYK theory simplifies significantly, and it becomes possible to exactly compute the disorder-averaged density of states using techniques from [62]. We sketch this in Appendix A.2.[22] To isolate the contribution of the Schwarzian, we take a further "triple-scaled" limit where we take $\frac{N}{q^2}$ large and the energy $(E - E_0)$ above the ground state small, with the product held fixed. In this limit, we find the density

[22] Notice that if we take a double scaling limit $q, N \to \infty$ keeping q/N^α fixed, then the scrambling time is of order $\log(N)$ when $\alpha < \frac{1}{2}$, while it is of order 1 when $\alpha > \frac{1}{2}$. Therefore $q^2 \sim N$ marks the boundary between the behaviors expected for k-local and nonlocal Hamiltonians [63–65].

of states (see Appendix A.2 eq. (A.20))

$$\rho(E) \propto \sinh\left(\pi\sqrt{2z}\right), \qquad z = \frac{(E-E_0)N}{q^2\mathcal{J}} \to \frac{4\alpha_S(E-E_0)N}{\mathcal{J}} = \frac{c(E-E_0)N}{\pi^2}. \tag{2.46}$$

We expect that the Schwarzian sector is the only part of the theory that survives this triple-scaled limit, so (2.46) should be an exact result for the Schwarzian theory. Computing the partition function via $Z = \int dE\rho(E)e^{-\beta E}$, we learn that the one-loop result (2.44) is actually the exact answer for the Schwarzian theory.

The conclusion of this discussion is that we can include the effect of the soft mode integral by simply dividing the large N saddle point expression for the partition function by a factor of $(\beta+it)^{3/2}$. Using the expression for the large N free energy in the holographic limit, $\log Z = N(\epsilon_0\beta + s_0 + \frac{c}{2\beta})$, one finds that the disconnected term in (2.42) contributes the following to $g(t)$:

$$\frac{|\langle Z(\beta+it)\rangle|^2}{\langle Z(\beta)\rangle^2} = \frac{\beta^3}{(\beta^2+t^2)^{3/2}}\exp\left(-\frac{cNt^2}{\beta(\beta^2+t^2)}\right). \tag{2.47}$$

The time dependence of the exponent becomes negligible at $t \gtrsim \sqrt{N}$, and we have a power law decay $\sim t^{-3}$. Away from the holographic limit, one would replace the piece in the exponential by the appropriate finite β saddle point action, which can be computed numerically.

Now, we would like to evaluate the second term in (2.42). Away from the holographic limit, one has to use the numerical $S(E)$ determined by solving the Schwinger-Dyson equations. However, we can give a simple formula in the holographic limit, where $S(E) = Ns_0 + \sqrt{2c(E-E_0)N}$. Neglecting one-loop factors

from the integral over E, we have the contribution to $g(t)$

$$g_{ramp}(t) \sim \begin{cases} \frac{t}{2\pi} \exp\left[-2Ns_0 - \frac{cN}{\beta}\right] , & \frac{t}{2\pi} < e^{Ns_0} \\ \frac{t}{2\pi} \exp\left[-2Ns_0 - \frac{cN}{\beta} - \frac{\beta}{cN} \log^2\left(\frac{t/(2\pi)}{e^{Ns_0}}\right)\right] , & e^{Ns_0} < \frac{t}{2\pi} < \frac{t_p}{2\pi} \\ \exp\left[-Ns_0 - \frac{3cN}{4\beta}\right] , & t_p < t. \end{cases} \quad (2.48)$$

where $t_p = 2\pi e^{Ns_0 + \frac{cN}{2\beta}} = 2\pi e^{S(\beta)}$. Notice that this function is C^1. We can evaluate the dip time by equating (2.47) and (2.48), which gives $t_d \sim e^{Ns_0/2}$.

One can also make a more exact analysis of the large N function, by evaluating the finite β saddle point action numerically, and doing the integral over E in (2.42). In Fig. 2.7 we show the result of doing this computation and plugging in $N = 34$ to compare to the exact diagonalization data. We also take into account the two-fold degeneracy in the spectrum for $N = 34$ and evaluate the numerical finite temperature saddle for the slope portion, slightly correcting (2.47). The agreement is reasonably good, although the ramp and plateau are off by factors that represent the discrepancy in the exact free energy vs. the large N saddle point. (Presumably this factor would be mostly resolved by a complete one-loop correction to the large N partition function.)

We caution the reader that although the agreement in Fig. 2.7 looks reasonable, it is very possible that the true large N answer for $g(t)$ would differ in important ways. In particular, we are not confident that the slope region continues to be described by the simple Schwarzian effective theory out to very long times of order $e^{Ns_0/2}$. Another possibility is that some effect leads to the slope portion of $g(t)$ decreasing more rapidly at an earlier timescale. For example, this could be the

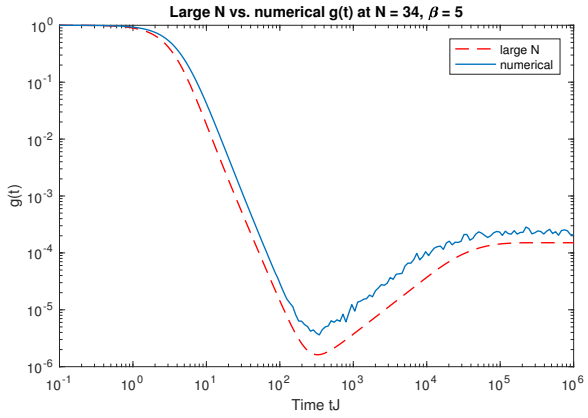


Figure 2.7: Comparision of (2.42) evaluated for the SYK model using the large N density of states extrapolated to $N = 34$ plus the Schwarzian partition function (red/dashed) against the $N = 34$ exact diagonalization answer for $g(t)$ (blue/solid). The discrepancies in the ramp and plateau regions are due to the fact that the large N free energy (without a proper one-loop term) gets the partition function wrong by an order one factor. In the ramp we are dividing by this twice, and in the plateau we are roughly dividing by it once.

result of some $1/N^q$ effect that tends to smooth out the sharp $\sqrt{E - E_0}$ edge in the spectrum, leading to a faster decay.[23] In this situation, the slope would crash and intersect the ramp much sooner, leading to a short dip time, perhaps of order $t_d \sim N^q$. Another possibility is that the very bottom of the spectrum would be controlled by a spin-glass phase that was argued to exist in the Sachdev-Ye model [52], and may also be present at exponentially low temperatures in the SYK model [66]. Such effects may also lead to a softer edge in the spectrum, again leading to a faster decay of the slope and a correspondingly shorter dip time.

We are fairly confident that the dip time should be no *later* than $e^{Ns_0/2}$, based on the idea that neglected effects are not likely to make the spectrum vanish more

[23] A simple example of a $1/N^q$ effect is the sample-to-sample variation of the edge of the eigenvalue spectrum. This causes a gaussian crash in the partition function $g_d(t)$ at times of order $N^{(q-2)/2}$ but cancels out in $g(t)$. Roughly speaking, effects that cause a crash in $g(t)$ must be present in a single sample.

sharply. As an extreme fallback position, one can argue without any calculation that the dip time is less than e^{Ns_0}, which is enough to establish a parametrically long ramp at non-zero temperature. To make the argument, one assumes that the slow decay in the slope is monotonic and roughly independent of temperature, based on the idea that it comes from the edge of the spectrum. Note that t_d can never be larger than t_p, because at times $t > t_p$ the spectral form factor $g(t)$ is only sensitive to individual energy levels, with all correlations between different levels getting washed out by the oscillating terms. For $t > t_p$, $g(t)$ is equal to the constant plateau height g_p. This allows us to conclude that $t_d(\beta) \leq t_d(\beta = \infty) \leq t_p(\beta = \infty) = e^{Ns_0}$.

2.7 Correlation functions

In this section we will discuss when two-point correlation functions exhibit ramp + plateau structure at late times. We will use the Eigenstate Thermalization Hypothesis (ETH) [67,68] to estimate matrix elements. As we will see, the answer depends on the (N mod 8) symmetry pattern [35, 36, 43], which is reviewed in Appendix A.1.

As before, we focus on the annealed ('factorized') two-point function

$$G(t) \equiv \frac{1}{N} \sum_{i=1}^{N} \frac{\langle \text{tr} \left[e^{-\beta H} \psi_i(t) \psi_i \right] \rangle_J}{\langle Z(\beta) \rangle_J}, \tag{2.49}$$

in which the disorder average is taken separately in the numerator and the denominator. This quantity is easier to study analytically than the quenched correlator. We note in passing that it is sometimes useful to consider the average of the

squared two-point function [10, 16], but for our purposes it will be enough to consider the average of the two-point function itself.

Let us first determine whether the two-point function has a nonzero plateau. This can be determined by considering the following long-time average in a single realization of the random couplings.

$$\frac{1}{Z(\beta)} \lim_{t_o \to \infty} \frac{1}{t_o} \int_0^{t_o} dt \, \mathrm{tr} \left[e^{-\beta H} \psi(t) \psi(0) \right] = \frac{1}{Z(\beta)} \sum_{\substack{n,m \\ E_n = E_m}} e^{-\beta E_n} \left| \langle n | \psi | m \rangle \right|^2 . \quad (2.50)$$

Here, $|n\rangle$ is the energy eigenbasis with energies E_n in the random couplings realization, and ψ stands for any one of the fermions ψ_i. (We neglect the effect of degeneracies for simplicity.) We expect a non-zero plateau to appear unless the matrix element vanishes. If $N/2$ is even then there is no degeneracy between the charge parity odd and even sectors (see Appendix A.1). In this case the matrix element in (2.50) equals zero and the plateau vanishes.

If $N/2$ is odd then we can use the particle-hole operator P to write down a selection rule for the matrix element. Let $|n\rangle$, $|m\rangle$ denote degenerate states with $E_n = E_m$, such that $P|m\rangle = |n\rangle$. Then we are interested in whether $\langle m|\psi|n\rangle$ can be nonzero. We have

$$\langle m|\psi|n\rangle = \Big(P\psi P|m\rangle, P|m\rangle \Big) = \eta(N) \Big(\psi|m\rangle, P|m\rangle \Big) = \eta(N) \langle m|\psi|n\rangle, \quad (2.51)$$

where we used inner product notation $\big(|1\rangle, |2\rangle\big) = \langle 1|2\rangle$ for clarity. In the second equality we used the antiunitarity of P, and in the third equality we used (A.3). We conclude that a plateau can only appear when $\eta(N) = 1$, or equivalently when

$(N \bmod 8) = 2$.

Next we ask when will a ramp appear in the two-point function. To answer this question, we put the disorder-averaged two-point function in the following form.

$$G(t) = \frac{1}{N\langle Z(\beta)\rangle_J} \sum_i \int dE dE' e^{-\beta E} e^{i(E-E')t} \left\langle \rho(E)\rho(E') \left|\langle E|\psi_i|E'\rangle\right|^2 \right\rangle_J . \quad (2.52)$$

Here, $\rho(E)$ is the energy spectrum in a given realization of the random couplings. Notice again that the matrix element $\langle E|\psi_i|E'\rangle$ connects eigenstates from two different charge parity sectors. We will again consider two separate cases, depending on whether $N/2$ is even or odd.

If $N/2$ is even then there is no degeneracy between the two sectors. The two ρ factors that appear in (2.52) are de-correlated for sufficiently small energy differences (corresponding to late times), and we do not expect a ramp to appear.

If $N/2$ is odd then the two charge parity sectors are degenerate, so effectively there is only one sector. As discussed above, at late times the correlator probes small energy differences in the spectrum, where we expect each sector of the Hamiltonian to resemble a Gaussian random matrix. For such a matrix, the averages over eigenvalues and eigenstates factorize, and we can approximate

$$\left\langle \rho(E)\rho(E') \left|\langle E|\psi_i|E'\rangle\right|^2 \right\rangle_J \approx \langle \rho(E)\rho(E')\rangle_J \cdot \left\langle \left|\langle E|\psi_i|E'\rangle\right|^2 \right\rangle_J . \quad (2.53)$$

Furthermore, for a Gaussian random matrix $\left\langle \left|\langle E|\psi_i|E'\rangle\right|^2 \right\rangle_J$ is a smooth function of the small energy difference $|E - E'|$, as in ETH, and we approximate it by a constant. The value of this function at $E = E'$ determines whether there is a non-

zero plateau, as discussed above. The remaining factor $\langle \rho(E)\rho(E')\rangle_J$ gives the spectral form factor. It will lead to a ramp, just as in the case of the observable $g(t)$ discussed in previous sections.

To summarize, the two-point function will display the following combinations of a ramp and a non-zero plateau, depending on the value of (N mod 8).

- If (N mod 8) $= 0, 4$ then there will be no ramp or plateau.
- If (N mod 8) $= 2$ then there will be a ramp and a non-zero plateau.
- If (N mod 8) $= 6$ then there will be a ramp but no plateau (the two-point function will vanish at late times).

Figures 2.8, 2.9 show a numerical computation of the two-point function that bears out these conclusions.

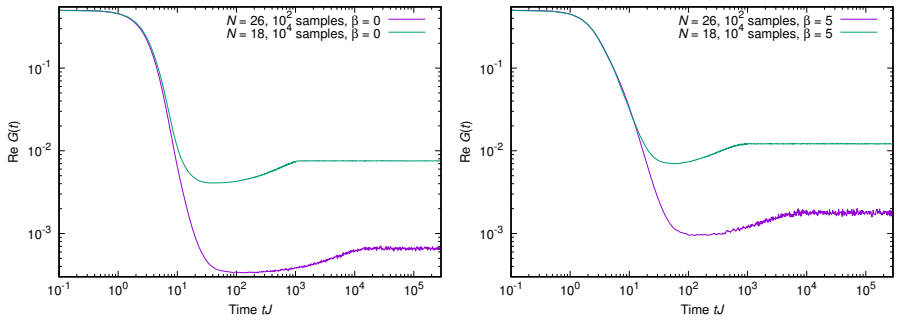


Figure 2.8: SYK two-point function (2.49) for $N = 18, 26$, plotted for $\beta = 0, 5$. A slope, dip, ramp, and plateau can be seen.

Finally, let us estimate the height of the correlator plateau written down in (2.50) in the cases where it does not vanish. This requires us to estimate the typical size of the matrix elements $|\psi_{nm}|^2 = |\langle n|\psi|m\rangle|^2$. For typical eigenstates $|n\rangle$ and $|m\rangle$, ETH predicts that the matrix elements will be of the same order

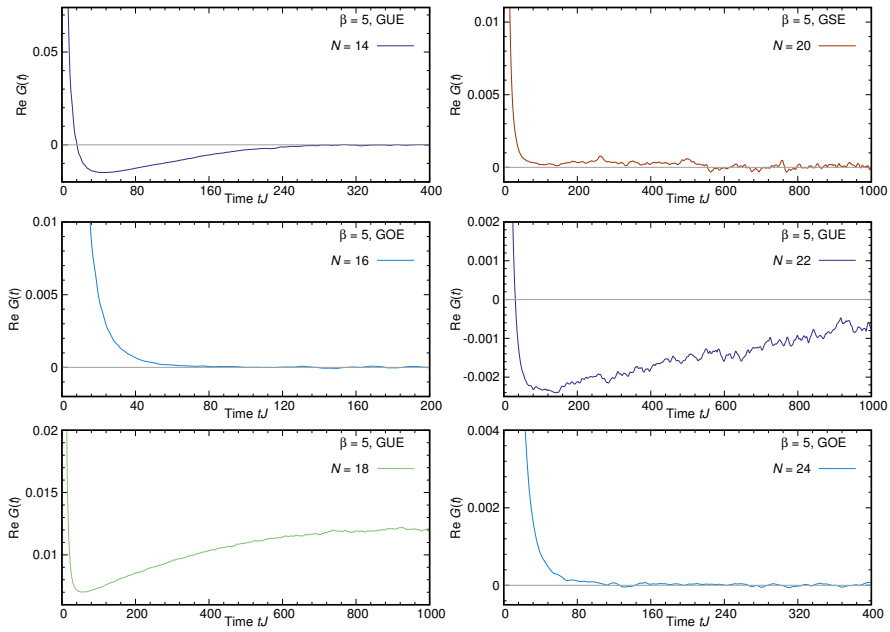


Figure 2.9: SYK two-point function $G(t)$ for $N = 14, 16, 18$ (left) and $N = 20, 22, 24$ (right), $\beta = 5$. The number of samples is 10^4 for $14 \leq N \leq 18$, 10^3 for $N = 20$, and 10^2 for $N = 22, 24$. A ramp appears for N mod $8 = 2, 6$ but not for N mod $8 = 0, 4$. A non-zero plateau appears only for N mod $8 = 2$. These properties are all explained by the (N mod 8) symmetry pattern.

as for random states, which would give $|\psi_{nm}|^2 = 1/L$. However, in our case, $|n\rangle$ and $|m\rangle$ are related by the action of the P operator, $|n\rangle = P|m\rangle$, so we are actually considering a diagonal expectation value $|\psi_{nm}|^2 = |\langle m|\psi P|m\rangle|^2$. Then ETH instructs us to estimate this by replacing $|m\rangle$ with a random state. One can check that this also gives $|\psi_{nm}|^2 \sim 1/L$. The height of the plateau in the correlation function should then be of order $1/L$.

Notice that the spectral form factor plateau at $\beta = 0$, given by (2.4), is also of order $1/L$. We therefore expect the correlator plateau and the spectral form factor plateau to be of the same order. This holds true for the $N = 18$ data, where the correlator plateau height is approximately 0.0075, and the spectral form factor

plateau is at approximately 0.004.

2.7.1 The ramp in more general theories

One goal of this work is to evaluate how generic the ramp/plateau structure is in chaotic quantum field theories. In this section we ask whether it is plausible for this structure to appear in two-point functions of the form $G(t) = \langle O(t)O(0)\rangle$ in such theories. We will make two assumptions: that ETH holds for the theory, and that the late time behavior includes a ramp, as predicted by RMT.

If O is a simple operator, we expect the two-point function to approach its time average G_p plus fluctuations of order $e^{-S} \sim 1/L$. We would like to make sure that ramp behavior is consistent with this expectation.

The correlator can be written in the energy basis as

$$\langle O(t)O(0)\rangle = \frac{1}{Z(\beta)}\sum_n e^{-\beta E_n}|O_{nn}|^2 + \frac{1}{Z(\beta)}\sum_{\substack{n,m \\ E_n\neq E_m}} e^{-\beta E_m}|O_{nm}|^2 e^{i(E_m-E_n)t}\,. \tag{2.54}$$

Here we assumed the spectrum is non-degenerate for simplicity. The first sum in (2.54), coming from terms with $E_n = E_m$ in the double energy sum, exactly gives the plateau height G_p. If the diagonal matrix elements $|O_{nn}|^2$ are of order unity then we find a plateau height $G_p \sim 1$ as discussed above.

The second sum in (2.54) encodes correlations between different energy levels. Beyond the dip time, it is responsible for the linear time dependence of the ramp. ETH predicts that the off-diagonal matrix elements $|O_{nm}|^2$ are of order $1/L$ — much smaller than the diagonal ones. To get an estimate for the second sum

we assume that these matrix elements can be treated as constant, $|O_{\text{off}-\text{diag.}}|^2 \sim 1/L$. The remaining sum then describes the ramp of the spectral form factor, sans the plateau contribution. Altogether, the two-point function (2.54) is given schematically by

$$\begin{aligned} G(t) &\sim G_p + |O_{\text{off}-\text{diag.}}|^2 \cdot Z(\beta) \cdot \left(\frac{t}{L^2} - \frac{1}{L}\right) \\ &\sim G_p + \frac{t}{L^2} - \frac{1}{L}\,. \end{aligned} \tag{2.55}$$

Note that the $Z(\beta)$ factor in front of the parentheses is needed because the correlator is normalized differently than the spectral form factor. In writing the above expression, we are imagining that we are averaging over time somewhat, in order to supress fluctuations of $G(t)$ and get a smooth ramp. This type of averaging will be discussed further in section 2.8. In any case, the conclusion of this analysis is that $\langle O(t)O(0)\rangle - G_p \sim \frac{t}{L^2} - \frac{1}{L}$; the difference is suppressed by $1/L$ as expected.

2.8 Single realization of random couplings

It is important to ask whether the late time features of the spectral form factor (the dip, the ramp and the plateau) appear in ordinary chaotic quantum field theories without a disorder average. As a first step towards addressing this question, in this section we consider the SYK spectral form factor $g(t;\beta)$ (2.10), computed for a single realization of the random couplings J_{ijkl}. Figure 2.10 compares the single sample result with the disorder averaged spectral form factor. Before the dip time there is good agreement between the single sample and the

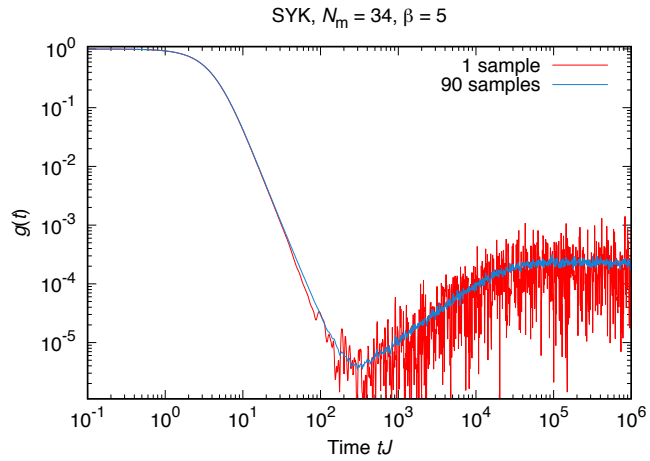


Figure 2.10: A single sample (red, erratic) of $g(t)$ is plotted together with the average of 90 samples (blue, smooth).

averaged results. This is consistent with our expectation that in the large N limit and at early times, a typical sample should give a good approximation to the disorder-averaged spectral form factor. We say that the spectral form factor is *self averaging* at early times.

At late times, and in particular after the dip time, the spectral form factor is not self averaging [69]. This implies that at large N a typical realization of the couplings does not give a result that approaches the disorder-averaged value. In particular, at late times a typical realization exhibits large fluctuations as shown in Figure 2.10. We expect ordinary quantum field theories (with no disorder average) to have similar behavior.[24]

Despite the large fluctuations, the underlying dip, ramp and plateau are still clearly visible. These features can be made clear by averaging over a sliding time window of width t_{ave}, smearing out the fluctuations.[25,26] For this to work we need

[24] We also expect the model recently discussed in [70] to behave similarly.

[25] Such time averaging and estimates compatible with ours have already been discussed in [69].

[26] Another possible way to reduce fluctuations in a CFT is to introduce a weak form of disorder

to be able to take $t_{\rm ave}$ parametrically shorter than the length of the ramp, so that the features we are interested in will not get smeared out along with the fluctuations. To estimate the required size of $t_{\rm ave}$, consider the auto-correlations in the random variable $|Z(\beta,t)|^2$,

$$h(t,dt;\beta) = \langle |Z(\beta,t)|^2 |Z(\beta,t+dt)|^2\rangle - \langle |Z(\beta,t)|^2\rangle \langle |Z(\beta,t+dt)|^2\rangle \,. \qquad (2.56)$$

Set t to be a fixed time greater than the dip time. At such fixed t the autocorrelation $h(t,dt;\beta)$ decays with dt with a typical time scale $t_{\rm decay}$. After $t_{\rm decay}$ the signal is essentially uncorrelated. As we will see shortly, for t on the plateau and at large N and large β we have $t_{\rm decay} \sim 1/\sqrt{N}$. For t on the ramp $t_{\rm decay} \sim 1$ (We have suppressed the β, J dependence here).

Given $t_{\rm decay}$ we can estimate the minimal value of $t_{\rm ave}$ required to remove the fluctuations from the single-sample data. We expect the fractional standard deviation (the ratio of the standard deviation to the mean) for such averaged data to behave like $\sqrt{\frac{t_{\rm decay}}{t_{\rm ave}}}$. To have a curve with fluctuations smaller than, say, $1/N^2$ would require $t_{\rm ave}$ to be no greater than N^4, even if $t_{\rm decay}$ is as large as 1. Such a value of $t_{\rm ave}$ is parametrically smaller than the length of the ramp, which is exponentially large in N. Therefore, at large N the averaging width $t_{\rm ave}$ can be taken to be parametrically smaller than the length of the ramp.

Now we show that on the plateau $t_{\rm decay} \sim 1/\sqrt{N}$ at large N and large fixed

averaging, by averaging slightly over the value of a marginal coupling.

β. In the energy eigenbasis the autocorrelation function can be written as

$$h(t, dt; \beta) = \sum_{k,l,m,n} \langle e^{-\beta(E_k+E_l+E_m+E_n)} e^{it(E_k-E_l)+i(t+dt)(E_m-E_n)} \rangle - \sum_{k,l,m,n} \langle e^{-\beta(E_k+E_l)} e^{it(E_k-E_l)} \rangle \langle e^{-\beta(E_m+E_n)} e^{i(t+dt)(E_m-E_n)} \rangle \,. \quad (2.57)$$

Let t be greater than the plateau time t_p. For such t, and when dt is small, the first sum is dominated by terms which obey $E_k - E_l + E_m - E_n = 0$. For a chaotic spectrum we generally expect only two solutions. One solution, $E_k = E_l$, $E_m = E_n$, cancels with the disconnected part. The second solution, $E_k = E_n$, $E_l = E_m$, gives the approximate answer

$$h(t, dt; \beta) \approx \sum_{m,n} \langle e^{-2\beta(E_n+E_m)} e^{idt(E_m-E_n)} \rangle = \langle Z(2\beta) \rangle^2 \cdot g(dt; 2\beta) \,. \quad (2.58)$$

(Here we assumed that there are no degeneracies for simplicity.) We find that at very late times t the time dependence of the autocorrelation is given by the spectral form factor $g(dt; 2\beta)$ at *early* times.

At large β and small dt, equation (2.47) provides a good approximation to the spectral form factor, which decays as $g(dt; 2\beta) \sim \exp\left(-cNdt^2/(2\beta)^3\right)$. The typical decay time scales as $t_{\text{decay}} \sim 1/\sqrt{N}$, as advertised above. After a time of order a few β the exponential decay is replaced by a $1/(dt)^3$ power law decay. By this time the spectral form factor (and hence the autocorrelation h) is already exponentially suppressed. A $1/(dt)^3$ power law is integrable so it does not alter our above estimate for the required time averaging window.

On the ramp the analysis is more subtle. First we make the plausible as-

sumption that the leading multipoint energy eigenvalue correlation functions at the exponentially small scales appropriate to the ramp are the same as the RMT correlators, up to $1/N^q$ corrections. In GUE these correlators factorize into sums of products of sine kernels [40]. Then we can use a procedure like that leading to (2.42) to show that for most of the ramp after a time dt of order 1 the auto-correlation function $h(dt)$ decays like $1/(dt)^4$. For the earliest part of the ramp, $t < e^{Ns_0}$, $h(dt) \sim 1/(dt)^3$. These power laws are integrable and so we estimate that on the ramp $t_{\rm decay} \sim 1$. This means that $t_{\rm ave}$ can be chosen to make the error smaller than any power of N and still leave exponentially many data points on the exponentially long ramp. Numerics are not conclusive here, but do show a systematic decrease of error after time averaging.

2.9 Conjecture about super-Yang-Mills

The above ideas make it possible to give a conjecture about the behavior of $g(t)$ and correlation functions $G(t)$ in the canonical AdS/CFT duality AdS_5/CFT_4 where CFT_4 is the $N = 4$ SU(N) super Yang-Mills theory on S^3. We will assume that the fine grained spectral statistics of this system are described by random matrix theory. This seems highly plausible given that this system at large 't Hooft coupling λ is maximally chaotic, i.e., it saturates the chaos bound [5]. We also assume that there are no new intervening nonperturbative time scales governing the behavior of $g(t)$ between the relatively short times governed by gravity and the very long times governed by random matrix theory. A distinctive aspect of this system compared to SYK is that at very high temperatures T the entropy

becomes parametrically large and the plateau parametrically high relative to the dip.

There is no ensemble of Hamiltonians in this system so we want to describe the time averaged behavior of $Z(\beta, t)$ as discussed in the previous section. We can relate this to the full density of states $\rho(E)$

$$Z(\beta, t) = \int_0^\infty dE \rho(E) e^{-(\beta+it)E} \ . \tag{2.59}$$

At early times and large N^2 we evaluate (2.59) by saddle point and use the bulk gravitational action to determine $\rho(E)$. The initial behavior of $Z(\beta + it)$ should then be given by analytically continuing the large euclidean AdS–Schwarzschild black hole action to complex β.

In the following we use the results and follow the notation of [71]. The black hole metric has warp factor $V(r) = 1 - \mu/r^{n-2} + r^2/l^2$ where $\mu \sim G_n M$, $n+1 = D$ is the bulk spacetime dimension, and l is the AdS radius. The horizon radius r_+ is determined by $V(r_+) = 0$.

The inverse temperature is determined by finding the periodicity of time of the Euclidean signature metric,

$$\beta = 4\pi(l^2 r_+)/(n r_+^2 + (n-2)l^2) \ . \tag{2.60}$$

The action I, $(Z = e^{-I})$ is given by

$$I_{BH} = \frac{C}{G_N}\beta\left(-r_+^n + r_+^{n-2}l^2 + \frac{3}{8}l^4\right) \ . \tag{2.61}$$

Here and below C is a positive constant and $G_N \sim \frac{1}{N^2}$. The $\frac{3}{8}l^4$ term is specific to $n = 4, D = 5$. These Casimir energy type terms are missed without thinking about holographic renormalization [71]. The thermal AdS action in this scheme is

$$I_{AdS} = \frac{C}{G_N}\beta \cdot \frac{3}{8}l^4 \,. \tag{2.62}$$

We find it convenient to subtract the ground state energy, and study (2.59) via e^{-I} where $I = I_{BH} - I_{AdS}$. In other words, we do not include the Casimir term.

Now we analytically continue. As $\beta \to \beta + it$, r_+ becomes complex. For small real β, $r_+ \sim 1/\beta$. Adding a small positive imaginary part to β corresponds to adding a small negative imaginary part to r_+. At large t, $r_+ \to -i\sqrt{\frac{n-2}{n}}l$.

At $t = 0$, $Z \sim e^{cN^2/\beta^3}$ which for small β is very large. This reflects the very large entropy at high temperatures. (Here and below c denotes positive constants of order one.) As t is increased and $\beta \to \beta + it$, $|Z|^2$ drops very quickly. At $t \sim \beta$, $|Z|^2$ becomes less than one. Then another saddle, the thermal AdS solution, dominates. Including the one loop determinant around this saddle representing a gas of gravitons we find $|Z|^2 \sim e^{c/\beta^3}$, an N-independent much smaller value.[27]

But this is not the whole story. As t increases to AdS scale $|Z|^2$ increases again, eventually becoming of order $|Z|^2 \sim e^{cN^2}$ (with no $\frac{1}{\beta^3}$ enhancement). This apparently dominates over the thermal AdS again.

But there is another wrinkle. As t becomes large the solutions of $V(r_+) = 0$ coalesce. This causes the fluctuation corrections to the saddle point in (2.59) to behave like $\frac{t}{N^2}$, becoming large at $t \sim N^2$. Taken at face value these large

[27] To be precise, there are of order $\frac{1}{\beta^3}$ weakly interacting gravitons of AdS energy and so $Z(\beta + it)$ oscillates with AdS frequency.

corrections invalidate the saddle point analysis for times larger than this.

Other saddle points could be relevant here. For example, at high temperature, small β, there is a 10D small Schwarzschild black hole saddle point with $r_+ \ll l$. Using the $n = 9$ version of the above formulas gives an initial $|Z|^2 \ll 1$, evolving at $t \sim \beta$ to $|Z|^2 \sim e^{cN^2}$ and then rapidly decreasing to $|Z|^2 \ll 1$ again. But at $t \sim l$, r_+ becomes of order l and AdS corrections become important. A more careful analysis would be required.

Although it is never thermodynamically dominant, the recent analysis of [72] indicates that there is a Gregory-Laflamme-type 5D to localized 10D transition in the space of saddles. At first glance this could produce a singularity in $\rho(E)$ leading to a slow long-time falloff. If this transition is caused by a single mode becoming tachyonic then it produces a branch point singularity in Z which presumably can be analytically continued around. If there is a more serious kind of large N transition it may produce a more extreme form of singularity. In any event, at large but finite N this feature will be smoothed out, so we do not expect it to produce significant long-time effects past times polynomial in N.

Although this analysis is far from conclusive[28] it does seem like the most plausible values of $|Z|^2$ in the slope region leading up to the dip have N scaling $|Z|^2 \sim 1$ or $|Z|^2 \sim e^{cN^2}$. We will assume these values and compute the dip by matching onto the ramp, to which we now turn.

[28] As an example of the subtleties here, for $n = 5, D = 6$ the dominant saddle causes $|Z|^2$ to diverge as $t \to \infty$. This is inconsistent so presumably this saddle eventually leaves the integration contour. In general we have not attempted to decide which saddles are on or off the steepest descent contour.

2.9.1 The ramp in SYM

SYM at large 't Hooft coupling λ is maximally chaotic according to the out of time order correlator diagnostic, so it is plausible to conjecture that its fine grained eigenvalue statistics are described by random matrix theory. The relevant ensemble will be determined by the symmetries of the system. For simplicity let us imagine that a nonzero θ term is present to break the T symmetry. Then we expect GUE statistics.[29] For simplicity, we will discuss the case of a high temperature state, where β is small compared to the spatial S^3 radius.

We can outline a simple expectation for the ramp behavior using the formula (2.42). We interpret the expectation value $\langle\cdot\rangle$ as denoting a time average rather than a disorder average, as discussed in section 2.8. The procedure is equivalent to "unfolding" the spectrum, analyzing the ramp and plateau for each narrow energy band, and then adding them up together. First, we study the ramp at reasonably late time, $t > e^{\#N^2}$, where the relevant energies will be high enough that we can use planar SYM formulas for $S(E)$:

$$\log Z = S - E/T = c_0 N^2 T^3 \,, \tag{2.63}$$

$$E = 3c_0 N^2 T^4 \,, \tag{2.64}$$

$$S = 4c_0 N^2 T^3 = \frac{4}{3^{3/4}} c_0^{1/4} N^{1/2} E^{3/4}. \tag{2.65}$$

At large N the integral in (2.42) will be sharply peaked, and the band that makes the largest contribution at time t will be the band which is just reaching the plateau at time t. That is, $S(E) = \log t$. Using the above equations we then

[29] We thank Alex Maloney for this observation.

have

$$g_{\rm ramp}(t) = \frac{t}{Z(\beta)^2} \exp\left[-\frac{3\beta}{2^{5/3}c_0^{1/3}} \frac{(\log t)^{4/3}}{N^{2/3}}\right]. \tag{2.66}$$

The growth is somewhat slower than linear, and the ramp joins the plateau at time $t_p = e^{S(2\beta)} = e^{4c_0N^2/(2\beta)^3}$ where the derivative of (2.66) vanishes.

To understand the dip time t_d we need to work out the behavior of the ramp at earlier times. It is possible that weak interactions in the AdS gas could lead to a small ramp, but we focus our attention on the region where the ramp would be associated to black hole states. The smallest black holes that dominate the microcanonical ensemble are determined by microscopic parameters, as discussed by [73–75], but in fact these black holes give contributions to the ramp that are smaller than the slope contribution to $g(t)$. To see this, we consider 10D Schwarzschild black holes of mass E with Schwarzschild radius r_s much smaller than the AdS radius l where (in the remainder of this section we suppress numerical factors)

$$E = r_s^7/G_N\,,$$
$$S = r_s^8/G_N\,.$$

Here $G_N = l^8/N^2$ is in 10D. The contribution of such black holes to the ramp would be

$$\frac{1}{Z(\beta)^2}\int dr_s\, e^{-N^2\beta r_s^7/l^8} \min\left\{t, e^{N^2 r_s^8/l^8}\right\} \approx \frac{t}{Z(\beta)^2} \exp\left[-N^{1/4}\frac{\beta}{l}(\log t)^{7/8}\right], \tag{2.67}$$

where we used that the integral is dominated by the value of r_s such that $t =$

$e^{N^2 r_s^8/l^8}$. The dip time is the first time such that this contribution is larger than the contribution of the slope. The expected slope contribution is no smaller than $\frac{1}{Z(\beta)^2}$. Eq. (2.67) first exceeds this value when the dominant value of r_s is $r_s = \beta$, or equivalently at a time

$$t_d = e^{N^2\beta^8/l^8}. \tag{2.68}$$

So we see that the first black holes that are relevant are small, with $r_s \sim \beta$, but not microscopic. The associated dip time is exponential, but with a parametrically small coefficient at high temperature. Note that this conclusion is rather sensitive to the long-time behavior of the slope, so this identification of the dip time is tentative. For example, if we have $g_{slope} = \frac{1}{Z(\beta)^2} e^{cN^2}$ instead, then we would expect $t_d \sim e^{c' N^2}$ for an order one c'. Either way, at high temperature we have a large hierarchy between the dip time and the plateau time $t_p \sim e^{4c_0 N^2/(2\beta)^3}$, leading to a parametrically long ramp.[30] The early time behavior of AdS_3/CFT_2 is under greater analytic control and has been analyzed in [15] . There is also an exponential hierarchy, although not as large, in this system.

It would be interesting to consider observables that probe the ramp during earlier times where microscopic black holes are relevant. One possibility would be to directly consider a microcanonical partition function that selects this part of the spectrum.

There is a subtlety in these estimates. SYM and many other theories have global symmetries (like the SO(6) R symmetry). We expect the spectrum within each sector to have chaotic RMT correlations, but the different sectors would be

[30] In fact here the hierarchy is more dramatic than in SYK because the plateau can be made arbitrarily high by increasing the black hole temperature.

essentially uncorrelated. We expect the number of thermodynamically significant sectors at fixed β to be at most polynomial in the entropy S. If we denote the separate sectors by indices a, b we can write $g(t) = \sum_{a,b} g_{ab}(t)$ where g_{ab} contains the sum over energies in the fixed sectors a, b. The diagonal terms in this sum contribute as usual to the ramp and plateau; the off-diagonal terms have vanishing contribution at late time and large S. So the overall heights of the ramp and plateau are suppressed by polynomials in S. This effect is subleading to the exponential effects we are interested in and so we ignore it. We have confirmed these ideas in the Dirac SYK model which contains a U(1) charge.

2.10 Discussion

In this paper we have argued that the late time behavior of horizon fluctuations in large AdS black holes is governed by random matrix dynamics. Our main tool was the SYK model, which we used as a simple model of a black hole, adequate for such qualitative questions.

Using numerical techniques we established random matrix behavior at late times. We were able to determine the early time behavior precisely in the double scaling limit. This enabled us to give a plausible estimate for the dip time by computing the intersection of these two curves.[31] The dip time is exponentially late, and the ramp region, controlled by long-range spectral rigidity, is exponentially long, stretching until the asymptotic plateau behavior sets in.

It will be useful to have analytic insight into the ramp behavior in the SYK

[31] As noted earlier there could be new phenomena at early times, like spin-glass behavior, or $1/N^q$ effects absent in the double scaling limit. It seems quite plausible these would cause the slope to decay faster and so make the dip time earlier.

model. In Appendix A.6 we make some preliminary remarks about the origin of the e^{-2S} scale of the height of the ramp in this model.

We used these ideas to formulate a conjecture about more general large AdS black holes, like those dual to 4D SYM theory. Here we rely on the widely accepted intuition that the fine grained structure of energy levels in chaotic systems is described by random matrix theory. We then estimated the time at which the ramp appears by making a provisional estimate of the analytically continued 5D AdS-Schwarzschild black hole metric. Again we find an exponential hierarchy of scales.[32] The early time behavior of AdS_3/CFT_2 is under greater analytic control and has been analyzed in [15] . There is also an exponential hierarchy in this system.

In all of these situations the dip time does not signal a new physical phenomenon[33] – it is just the time when the ramp becomes larger in size than the slope. To understand the new physics of the ramp it would be interesting to follow it "underneath" the slope to see what happens at early times. For instance in SYM one would start accessing regions controlled by string scale black holes, and eventually the chaotic graviton gas. To do this it might be useful to use a more refined probe than $g(t)$.

A more indirect strategy would be to look for precursors of the ramp starting from short times. Do the individual terms in the $1/N$ expansion get large as time is increased?[34] Or is there just a factorial growth of coefficients signaling an

[32] In fact here the hierarchy is more dramatic because the plateau can be made arbitrarily high by increasing the black hole temperature.

[33] If the spin glass or $1/N^q$ possibilities are present then there is a new physical phenomenon at the dip.

[34] The disconnected partition function $g_d(t)$ provides an example of this. As discussed above, the gaussian fluctuations of the edge of the eigenvalue distribution produce a gaussian falloff in

asymptotic expansion with an exponentially small error sufficient to accommodate the ramp and plateau signals? Knowing this would be helpful in looking for signals in SYM of these phenomena.

To understand the SYM situation better it would be useful to understand more about the averaging procedures that are available. Averaging over time windows has been discussed in Section 2.8. But perhaps one could take an ensemble of SYM theories with slightly different parameters. This possibility may be easier to implement in calculations.

Another set of ideas that might be useful are developments in the theory of sparse random matrices. From this point of view the SYK model is a certain type of sparse random matrix with correlated randomness in the entries. Insights have emerged [62,76–78] about the universality of dense random matrix behavior in the fine grained eigenvalue statistics of various types of sparse matrices. These might give clues about the SYK model, and more general contexts. This is a question we would like to return to in future research.

Perhaps the central question this work raises is the nature of the bulk interpretation of the random matrix behavior. The disorder averaged SYK model can be rewritten exactly in terms of the bilocal collective fields $G(t,t'), \Sigma(t,t')$. For $g(t)$ one needs two copies ("replicas") of the fermions and so the $G_{\alpha\beta}, \Sigma_{\alpha\beta}$ fields carry replica indices $\alpha, \beta = 1, 2$. An appropriate contour can be chosen so the functional integral over $G_{\alpha\beta}(t,t'), \Sigma_{\alpha\beta}(t,t')$ is nonperturbatively well-defined, as discussed in Appendix A.3. This functional integral is a rough proxy for a bulk description, because it involves $O(N)$ singlet objects and in some rough way bulk

$g_d(t)$. These are signaled by a series of terms of the form $(t^2/N^{q-2})^k$. This softening cancels out in the time dependence of $g(t)$.

fields should be able to be reconstructed from the bilocal singlet fields.

This functional integral must contain the ramp and plateau behavior – the question is how. We cannot yet answer this question – it will continue to be a focus of our research. Here we will just make some preliminary comments.

The coefficient of the G, Σ action includes N, so new saddle points are a natural mechanism for such e^{-N} effects. For $q = 2$, as explained in Appendix A.5, quenched correlators do seem to be described by sums over new saddle points with appropriate fluctuation corrections. Here the ramp is a perturbative $1/N^2$ effect and the plateau is an e^{-N} effect.

For the interacting case $q > 2$ the situation is qualitatively different. Here the interplay between L and N discussed in Section 2.5 becomes crucial. The ramp is a $1/L^2$ effect, which means an e^{-N} effect. In Appendix A.7 we point out obstacles to a possible single saddle point explanation of the ramp in the correlator $G(t)$. But various auxiliary quantities like the f_k discussed in Appendix A.6 can be computed by saddle point, giving the desired 2^{-N} value for large k. It is unclear whether this has anything to do with an actual saddle point description of the ramp involving a sum over many saddle points.

The N mod 8 "eightfold way" pattern noted in [35, 36, 43] must have an explanation in the G, Σ integral. In some ways it seems analogous to the behavior of the Haldane spin chain [79] as the spin varies from integer to half integer. There the explanation in the continuum is a topological term in the action. That would be a natural guess here, and the question is what topology is being probed. As an initial step it will be important to find the origin of this effect in the moment calculations discussed in Appendix A.6.

The origin of the plateau in the G, Σ integral is another mystery. After continuing to imaginary energy this is an effect of order $\exp\{(-L)\}$ which is of order $\exp\{(-e^N)\}$. This is an unusually small nonperturbative effect, the size of the error in an asymptotic series of multi-instanton corrections. A more natural way to explain these effects would be a mapping from G, Σ to new effective random matrix degrees of freedom with effective coupling $1/L$ whose dynamics would give the plateau as a standard Andreev-Altshuler instanton nonperturbative effect [59, 60]. This map would be related to reconstituting the fermions from the collective fields.[35] This is a challenging proposition but the SYK model provides the most concrete arena known in which to explore it.

[35] Some ways in which fermionic properties are coded into G, Σ are discussed in Appendix A.3.

Chapter 3

Infinite Temperature Operator Growth

We discuss the probability distribution for the "size" of a time-evolving operator in the SYK model. Scrambling is related to the fact that as time passes, the distribution shifts towards larger operators. Initially, the rate is exponential and determined by the infinite-temperature chaos exponent. We evaluate the size distribution numerically for $N = 30$, and show how to compute it in the large-N theory using the dressed fermion propagator. We then evaluate the distribution explicitly at leading nontrivial order in the large-q expansion.

3.1 Introduction

In quantum many-body systems, the butterfly effect is roughly the statement that time evolution takes simple (few-body) operators to complicated ones (many-body). This makes it possible for the disturbance of a single particle far in the past to have significant effects on all particles at a later time. In systems with spatial locality, this takes a while, since the disturbance has to spread through the

system [27, 80–90]. In nonlocal systems the process can be much faster. However, the concept of operator growth still makes sense if each term in the Hamiltonian only couples together a few degrees of freedom at a time. In this setting, simple operators still take time to become complicated [63, 64, 91–93].

A rough diagnostic of this effect is the commutator-squared between $W(t) = e^{iHt}We^{-iHt}$ and V, where W, V are simple operators [21, 94–97]. The idea is that as time advances, $W(t)$ grows in such a way that it has nontrivial effects at almost any site in the system. As a result, it then fails to commute with other simple operators, such as V, and so $\langle [W(t), V]^2 \rangle$ becomes order one. In the case where W, V are fermionic operators, then one considers the anticommutator-squared instead.

In this paper we will consider another diagnostic, which is to compute the full probability distribution for the size of the time-evolving operator [82]. To define this, one expresses $W(t)$ in a basis of operators organized by the number of "simple" operators that appear in a given product (the "size"). Let's explain this more concretely for the case of the SYK model [4, 98]. In that case it is natural to take the simple operators to be the individual fermions, ψ_i. We choose W to be one of those fermions, say $W = \psi_1$. The time-evolving $W(t)$ can be expanded as

$$\psi_1(t) = \sum_{s,\, a_1 < \ldots < a_s} 2^{\frac{s-1}{2}} c_{a_1 \ldots a_s}(t) \psi_{a_1} \ldots \psi_{a_s}, \tag{3.1}$$

where s is the "size" of the basis element, i.e. the number of elementary fermions that appear in the product. The factor $2^{\frac{s-1}{2}}$ is to compensate for the fact that we normalize the fermions so that $\psi^2 = \frac{1}{2}$. The probability distribution for size s is

then defined as

$$P_s(t) = \sum_{a_1<\ldots<a_s} |c_{a_1\ldots a_s}(t)|^2. \tag{3.2}$$

As time passes, this distribution shifts towards operators of larger size—the operator grows.

We can think of the $c_{a_1\ldots a_s}(t)$ coefficients as describing a quantum wave function for the evolving operator. As we will see, in the infinite N SYK model, this "operator wave function" can be understood as a standard wave function for a quantum particle moving on a special, rapidly expanding graph shown in Fig. 3.2a. With time evolution, most of the particle's wave function moves deeper into the graph at an exponentially growing rate. This corresponds to the operator becoming larger and more complicated.

The fact that the rate is exponential is because the graph on which the particle is moving becomes more highly connected as we move deeper. In terms of the growing operator, this reflects the fact that once an operator has already become quite large, it has many different ways to grow larger still. This is the basic origin of exponential early-time behavior of correlators such as $\langle [W(t), V]^2 \rangle$, diagnosed by the chaos exponent (or many-body quantum Lyapunov exponent) λ_L [5, 97].

Although most of the wave function moves rapidly into the graph, there is a small exponentially decaying tail for the particle to remain at (or return to) the root of the graph. In operator language this is the amplitude for $\psi_1(t)$ to remain equal to $\psi_1(0) = \psi_1$, or more explicitly $\frac{1}{2^{N/2}}\mathrm{Tr}[\psi_1(t)\psi_1(0)]$, the infinite temperature two point function. This correlator exponentially decays because most of the wave function is leaking into the space of complicated many-fermion operators.

It is a challenging problem to go beyond this qualitative discussion and actually

compute $P_s(t)$. In the rest of the paper, we will discuss some partial results for this quantity, mainly in the large N theory. In particular, we will explain an equivalent particle-moving-on-a-graph problem, and we will show how to sum the 'melonic' (infinite N) perturbation theory for the wave function using the dressed infinite-temperature two-point function of fermions. However, the dressed two-point function is not known analytically except for the large-q SYK model. So the only place where we will succeed in computing the operator wave function is at infinite N and at leading order in large q. We will use this wave function to compute a few things, including the (previously known) infinite-temperature chaos exponent.

Before we begin with the main calculation, we will make three preliminary comments.

3.2 Preliminary comments

3.2.1 A classical model

Before analyzing the quantum problem, we can discuss an analog of operator growth for a classical model of many-body chaos. This model was previously considered in the context of high energy scattering in weakly coupled gauge theory [99]. It goes like this. Suppose we have a collection of N particles where we initially label one as infected and the others as healthy. The rule for time evolution is that with some probability γ per unit time, any infected particle can heal itself at the cost of infecting $(q-1)$ random other particles. For simplicity, we assume that the total number N is large enough and the time is short enough that the infected

particles are always very dilute. If we are interested in the probability that some randomly chosen particle will be infected after time t, we can proceed in three different ways. The three ways get increasingly complicated, but each gives an interesting perspective.

The first way is to simply notice that the expected number of infected particles $\langle N_{inf} \rangle$ is growing according to $\frac{d\langle N_{inf} \rangle}{dt} = (q-2)\gamma \ \langle N_{inf} \rangle$. This leads to $\langle N_{inf} \rangle = e^{(q-2)\gamma t}$. The probability that a randomly chosen particle will be infected at time t is simply $\frac{\langle N_{inf} \rangle}{N} = \frac{1}{N} e^{(q-2)\gamma t}$. This type of intuition was used in early discussions of scrambling by quantum circuits [63], and it is related to the kinetic equation method used for the OTOC in weakly coupled theories in [83].

A second way is to follow the possible chains of events that lead to our random final particle getting infected, and add up all of the probabilities. To start, the simplest way it could happen is that the original infected particle never infects any other particles, but by chance it happens to be our randomly selected particle at time t. The probability for this is $\frac{1}{N} e^{-\gamma t}$. The second simplest thing would be for the original particle to infect one set of $(q-1)$ particles, for one of these to be our randomly chosen particle, and for this one to not infect any further particles. The probability for this is

$$\frac{q-1}{N} \gamma \int_0^t dt' e^{-\gamma t'} e^{-\gamma(t-t')} = \frac{q-1}{N} \gamma t \, e^{-\gamma t}. \tag{3.3}$$

Summing over all possible chains of infection, we find the probability

$$\frac{1}{N} \sum_{k=0}^{\infty} \frac{\big((q-1)\,\gamma t\big)^k}{k!} e^{-\gamma t} = \frac{1}{N} e^{(q-2)\gamma t}. \tag{3.4}$$

This method of calculation is very similar to the calculation of the OTOC by ladder diagrams, where k labels the number of rungs in the ladder (see [19, 98] in the SYK model and [23, 85–87, 100] in weakly coupled theories). In particular the crucial factors of $(q-1)$ that appear here and that also appear in the ladder diagrams. These factors correct for the fact that we are only following one possible chain of infection, picking one out of the $(q-1)$ particles infected in each given event.

A final way, analogous to the discussion of the quantum problem in this paper, is to calculate the probability distribution for the number of infected particles at time t, and then take the expectation value explicitly. To do this we can use the equation

$$\frac{dP(s_k,t)}{dt} = \gamma s_{k-1}\, P(s_{k-1},t) - \gamma s_k\, P(s_k,t), \qquad s_k = 1+(q-2)k. \tag{3.5}$$

The solution with initial conditions $P(s_k,0)=\delta_{k,0}$ is

$$P(s_k,t) = \frac{\Gamma(k+\frac{1}{q-2})}{\Gamma(k+1)\Gamma(\frac{1}{q-2})} e^{-\gamma t}\big(1-e^{-(q-2)\gamma t}\big)^k. \tag{3.6}$$

With this probability distribution, we can of course find the same answer as with the other two methods for the probability that a random particle is infected, by taking the sum $\sum_k \frac{s_k}{N} P(s_k,t) = \frac{1}{N} e^{(q-2)\gamma t}$. It is interesting to note that if we scale time so that $\gamma \sim \frac{1}{q}$, this formula is quite similar to the probablity distribution we will find in the SYK model at large q, see (3.51).

3.2.2 A numerical plot for $N = 30$

We would now like to show an example numerical plot of $P_s(t)$ for the quantum problem. To begin we should explain how this can be computed numerically. The SYK model with N Majorana fermions lives in a Hilbert space of dimension $2^{N/2}$. The space of operators acting on that Hilbert space can be understood as a Hilbert space in its own right, with inner product given by

$$(A, B) \equiv \frac{1}{2^{N/2}} \text{Tr}\left[A^\dagger B\right]. \tag{3.7}$$

In this "operator Hilbert space" we can decompose $\psi_1(t)$ in a basis of operators of definite size, as in (3.1). A formula equivalent to (3.2) is

$$P_s(t) \equiv \sum_{\mathcal{O} \in \{\text{op.s of size } s\}} \frac{\left|(\mathcal{O}, \psi_1(t))\right|^2}{(\mathcal{O}, \mathcal{O})\,(\psi_1, \psi_1)}. \tag{3.8}$$

Here, the sum is over an orthogonal basis s-fermion operators, for example all $\binom{N}{s}$ operators of the form $\psi_{a_1} \dots \psi_{a_s}$ with $a_1 < ... < a_s$.

To evaluate this in practice for reasonably small values of N, we can compute $\psi_1(t)$ by exact diagonalization and exponentiation of the Hamiltonian and then evaluate the sum over operators in (3.8) by random sampling. In Fig. 3.1 we show the result of this computation. We also plot the expected value and variance of the size as a function of time. At $t = 0$ all of the probability is concentrated in size one, but as time passes we see successive peaks in larger sizes as the probability mass moves towards larger operators. At late time, the distribution appears to converge to the size distribution of a random fermionic operator. For such an

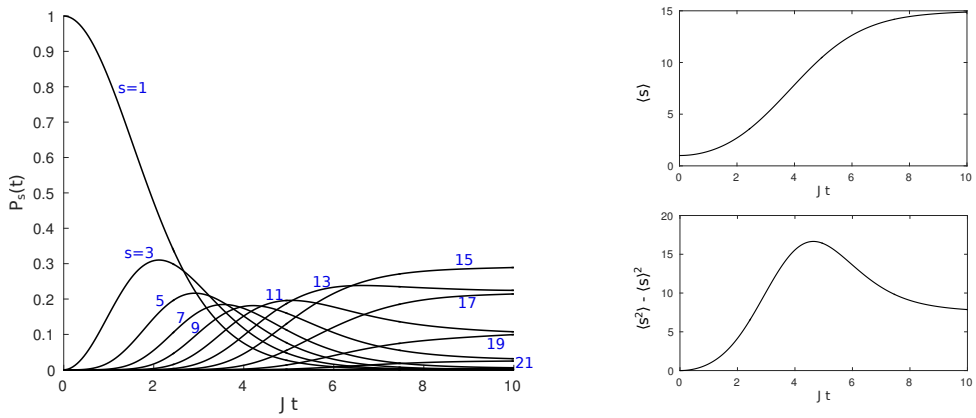


Figure 3.1: Exact diagonalization numerics for SYK with $N = 30$ and $q = 4$ (see § 3.3 for the definition of this model). At left, we plot the distribution of sizes in the operator $\psi_1(t)$ as a function of time. Notice that in the early phase, the peaks occur more rapidly as time passes. This is because already-large operators can grow faster than small ones. The "scrambling time" where the operator reaches full size would fall somewhere around three-quarters of the way through the plot. At right, we plot both the mean value and the variance of the size.

operator, P_s is proportional to the total number of operators of size s (for s odd only), which gives $P_s \to \binom{N}{s} 2^{1-N}$. So for example, the most common size at late time is $\frac{N}{2} = 15$.

3.2.3 The size and infinite-temperature OTOCs

As a final preliminary comment, we would like to show that out-of-time-order correlators at infinite temperature are related to the expectation value of s in the distribution $P_s(t)$. We define $A(t)$ as the typical anticommutator squared at infinite temperature between $\psi_1(t)$ and a single-fermion operator

$$A(t) \equiv \frac{1}{N} \sum_j \frac{1}{2^{N/2}} \mathrm{Tr}\left[\{\psi_1(t), \psi_j\}^2\right] = \frac{1}{N} \sum_j \Big(\{\psi_1(t), \psi_j\}, \{\psi_1(t), \psi_j\}\Big). \quad (3.9)$$

Here, we have averaged over the index of the second operator ψ_j. The growth of this object is a useful diagnostic for quantum chaos and is simply related to other infinite-temperature out-of-time-order correlators.

Inserting the expansion of $\psi_1(t)$ in (3.1), we have

$$A(t) = \frac{1}{N}\sum_j \sum_{\substack{s,\, a_1<\ldots<a_s \\ s',\, b_1<\ldots<b_{s'}}} 2^{\frac{s+s'-2}{2}} c^*_{a_1\ldots a_s}(t)\, c_{b_1\ldots b_{s'}}(t) \Big(\{\psi_{a_1}\ldots\psi_{a_s},\psi_j\},\{\psi_{b_1}\ldots\psi_{b_{s'}},\psi_j\}\Big). \tag{3.10}$$

In order to simplify this expression, we can use that the $\psi_{a_1}\ldots\psi_{a_s}$ operators are orthogonal with respect to our inner product $(\cdot,\cdot)$ and that this is preserved after taking anticommutators with ψ_j. In fact, when $a_1 < \ldots < a_s$ and $b_1 < \ldots < b_{s'}$, one has the useful formula

$$\Big(\{\psi_{a_1}\ldots\psi_{a_s},\psi_j\},\{\psi_{b_1}\ldots\psi_{b_{s'}},\psi_j\}\Big) = \begin{cases} 2^{1-s}, & \{a_1\ldots a_s\} = \{b_1\ldots b_{s'}\},\ j\in\{a_1\ldots a_s\}, \\ 0, & \text{else.} \end{cases} \tag{3.11}$$

This formula collapses the sum over $\{a_1\ldots a_s\}$ and $\{b_1\ldots b_{s'}\}$ to the diagonal terms. It also allows us to sum over j, getting a factor of s from the s different values of j for which we get a nonzero answer. We find

$$A(t) = \frac{1}{N}\sum_{s,\, a_1<\ldots<a_s} s\,|c_{a_1\ldots a_s}(t)|^2 = \frac{1}{N}\sum_s s\,P_s(t) = \frac{\langle s\rangle}{N}. \tag{3.12}$$

In other words, $A(t)$ is simply related to the mean size of the operator $\psi_1(t)$.

3.3 The graph of operators

We will now proceed with the main part of the paper. From this point forward we will be discussing the SYK model in the large N limit. Our conventions for SYK [4, 98] are that the Hamiltonian is

$$H = i^{q/2} \sum_{1 \leq a_1 < \ldots < a_q \leq N} J_{a_1 \ldots a_q} \psi_{a_1} \ldots \psi_{a_q}, \qquad \{\psi_a, \psi_b\} = \delta_{ab}. \tag{3.13}$$

Here $J_{a_1 \ldots a_q}$ is an antisymmetric tensor, drawn from a Gaussian distribution, with mean zero and the property that the square of a given component has the average value

$$\langle J^2_{a_1 \ldots a_q} \rangle = \frac{(q-1)!}{N^{q-1}} J^2, \qquad \text{(no sum)}. \tag{3.14}$$

Here, we introduced the dimensionful constant J. We will also use $\mathcal{J}$, which differs by a q-dependent factor as in [2]

$$J^2 = \frac{2^{q-1}}{q} \mathcal{J}^2. \tag{3.15}$$

We would like to understand the time evolution of a particular fermion operator $\psi_1(t)$ in the large N limit of this model. A key simplification will be that the this time evolution stays within a particular class of operators, which consist of many fermions contracted together in various ways with the $J_{a_1 \ldots a_q}$ tensor.

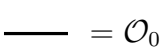
It is convenient to organize this class of operators by their size, which as always refers to the number of elementary fermion operators. We will sometimes use "generation" in place of size, where generation refers to the number of times we have to commute H with ψ_1 for the operator to first appear in the Baker-

Campbell-Hausdorff series for $\psi_1(t)$ (for further discussion of this perspective, see § 4.6). Size s and generation k are related by

$$s = 1 + (q-2)k, \qquad s = \text{size}, \qquad k = \text{generation}. \tag{3.16}$$

Let us now discuss the types of operators that appear in the time evolution of $\psi_1(t)$, choosing $q = 4$ for simplicity.

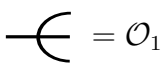
- **Generation zero:** At infinite N, the only size-one operator that appears in $\psi_1(t)$ is simply ψ_1 itself. It will be convenient to work with operators that are orthonormal with respect to the inner product defined in (3.7). A normalized version of the operator ψ_1 is simply $\sqrt{2}\psi_1$, which we denote as



$$= \mathcal{O}_0 = 2^{\frac{1}{2}}\psi_1. \tag{3.17}$$

  Our notation for this operator as a horizontal line will become clear from further examples below.

- **Generation one:** At generation one (size $q-1$), the operator that appears in the time evolution is simply the commutator of the Hamiltonian with ψ_1. The normalized version of this operator is



$$= \mathcal{O}_1 = 2^{\frac{3}{2}} \sum_{a<b<c} \frac{J_{1abc}}{J} \psi_a \psi_b \psi_c. \tag{3.18}$$

  We can interpret this operator as follows. The original fermion ψ_1 has split into $q-1$ fermions by a single action of the Hamiltonian.

- **Generation two:** In the second generation, it will be convenient to divide the operator into three (more generally $q-1$) terms, corresponding to a further division into $q-1$ fermions of any of the fermions present in the operator O_1. These distinct terms correspond to the operators

$$= \mathcal{O}_2^{(1)} = 2^{\frac{5}{2}} \sum_{\substack{a_1<a_2<a_3 \\ b_1<b_2<b_3}} \frac{J_{1a_1a_2a_3}J_{a_1b_1b_2b_3}}{J^2} \psi_{a_2}\psi_{a_3}\psi_{b_1}\psi_{b_2}\psi_{b_3}, \tag{3.19}$$

$$\vdots$$

$$= \mathcal{O}_2^{(3)} = 2^{\frac{5}{2}} \sum_{\substack{a_1<a_2<a_3 \\ b_1<b_2<b_3}} \frac{J_{1a_1a_2a_3}J_{a_3b_1b_2b_3}}{J^2} \psi_{a_1}\psi_{a_2}\psi_{b_1}\psi_{b_2}\psi_{b_3}. \tag{3.20}$$

Our notation with the fan diagrams is that the three daughter lines coming out of a vertex are always ordered such that the index of the top line is less than the index of the middle line, which is less than the index of the bottom line. Because of this ordering convention, the operators shown above are different from each other.

- **Generation three:** In the third generation, there are different kind of operators that can appear, corresponding to the division of a fermion that was "born" in the first generation or the second generation. For example, we have the operators

$$\ldots \qquad \ldots \tag{3.21}$$

- **Generation $k+1$:** More generally, the operators for generation $k+1$ are obtained by considering all of the operators at generation k, and for each

one, allowing one of the fermions to divide further, contracting with a $J_{a_1 \ldots a_q}$ symbol, and normalizing. Graphically, we simply turn one of the lines into a fan.

In the infinite N limit, these operators all have definite size and are orthogonal. Note that at finite N, some of the indices of the fermions might happen to be the same. Using that $\psi_a^2 = \frac{1}{2}$, this would imply that the operator is actually of smaller size than $1 + (q-2)k$. However, this does not happen at infinite N.

Now that we have discussed the set of operators that we will use, we can describe the evolution of $\psi_1(t)$. The idea is that the operators we have described form a graph, and the time evolution of the operator is simply the quantum evolution of a particle moving on the graph. More precisely, we can think of the space of operators being a Hilbert space with inner product (3.7). In the infinite N SYK model, the operators we dissused above correspond to an orthonormal basis for a subspace of the space of all possible operators. It is helpful to think about these operators $\mathcal{O}_k^{(\ell)}$ as basis states $|\mathcal{O}_k^{(\ell)}\rangle$ for an abstract particle that represents the evolving operator $\psi_1(t)$. The Heisenberg equation $\frac{d}{dt}\psi_1(t) = i[H, \psi_1(t)]$ is an ordinary Schrodinger equation acting in this space, for an appropriate Hamiltonian $\widehat{H}$:

$$\frac{d}{dt}|\psi_1(t)\rangle = -i\widehat{H}|\psi_1(t)\rangle, \qquad \langle A|\widehat{H}|B\rangle \equiv -\frac{1}{2^{N/2}}\mathrm{Tr}\left(A^\dagger[H,B]\right). \tag{3.22}$$

We can now explain the point of the basis of operators that we have chosen. The nice feature is that in this basis, $\widehat{H}$ is proportional to the adjacency matrix

on the graph,

$$\widehat{H} = 2^{1-\frac{q}{2}} J \cdot (\text{adjacency matrix}). \tag{3.23}$$

The adjacency matrix is defined as the matrix that has a 1 at location i, j if i, j are vertices connected by an edge, and a zero otherwise. So, for example, we have

$$\langle \text{—} | \widehat{H} | \text{—}\!\in \rangle = 2^{1-\frac{q}{2}} J, \qquad \langle \text{—}\!\in | \widehat{H} | \text{—}\!\in\!\in \rangle = 2^{1-\frac{q}{2}} J, \qquad \langle \text{—} | \widehat{H} | \text{—}\!\in\!\in \rangle = 0. \tag{3.24}$$

The evolution of the operator $\psi_1(t)$ in the large N theory is therefore simply the quantum evolution of a particle moving on the graph shown in Fig. 3.2, with initial condition that the particle starts out at the leftmost vertex.

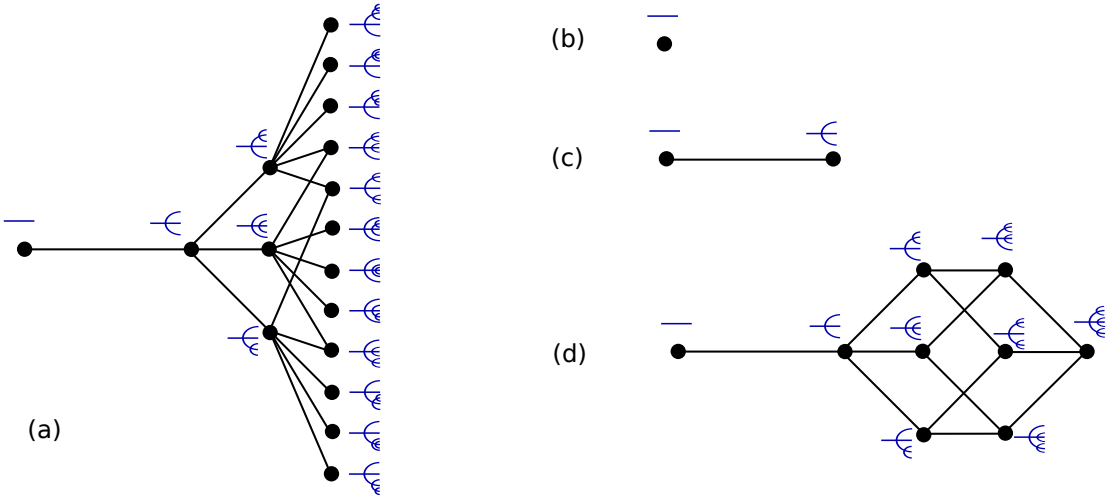


Figure 3.2: The graph of operators. In (a) we show the first four layers. Vertices correspond to basis operators, whose associated fan diagrams are indicated in blue. The problem of the time evolution of $\psi_1(t)$ in the large N theory is equivalent to the motion of a quantum particle on this graph (extended to further layers). In (b), (c), and (d), we show versions of the graph where we limit the recursive depth of the fan diagrams. The return amplitude on these graphs gives the zeroth, first and second iterations of the real-time Schwinger-Dyson equations. For any finite cutoff these amplitudes oscillate in time, but for the infinite graph the return amplitude decays exponentially.

An obvious feature of this graph is that it is rapidly expanding. The degree of a vertex (the number of neighbors) grows roughly linearly with generation k. This corresponds to the fact that at generation k, the operators contain $1 + (q-2)k$ fermions, and the Hamiltonian can act on any one of those, turning a single fermion into a fan of fermions and producing a new operator at generation $k+1$. The fact that the degree is growing with distance in this way means that this graph expands more rapidly than e.g. a Cayley graph/Bethe lattice/discretization of hyperbolic space, for which the degree is constant.

We would like to call attention to two qualitative features of the evolution of a particle on such a graph.

1. A basic effect is that particles tend to move to the right, towards more complicated operators. This is because at any given vertex, there tend to be many more edges leading to the right than to the left: there are more ways for the operator to grow than to shrink. We expect this to lead to exponential decay of the amplitude that the particle remains (or returns) to the original leftmost vertex ψ_1. This amplitude is simply the correlation function $\frac{2}{2^{N/2}}\mathrm{Tr}\left(\psi_1(t)\psi_1(0)\right)$. It exponentially decays due to the wave function for $\psi_1(t)$ leaking more and more into the space of complicated operators orthogonal to $\psi_1(0) = \psi_1$.

   The graph picture gives an intuitive explanation for why the real-time correlator should exponentially decay, but it does not give an efficient method for computing the decay rate. The best way we know of to compute the correlator is by numerically solving the Schwinger Dyson equations in real time to compute the retarded propagator. At infinite temperature, the real-time

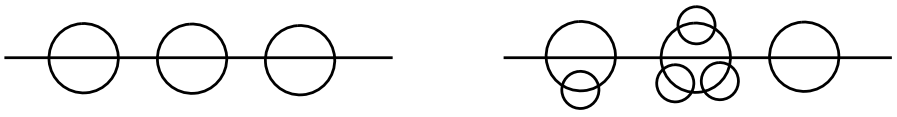
equations are simply given by

$$G_R(\omega) = \frac{1}{-i\omega + \Sigma(\omega) + \epsilon}, \qquad \Sigma(t) = 2^{2-q} J^2 G_R(t)^{q-1} \tag{3.25}$$

At infinite temperature, the retarded propagator is simply

$$G_R(t) = \frac{2\theta(t)}{2^{N/2}} \mathrm{Tr}\left(\psi_1(t)\psi_1(0)\right).$$

For $t > 0$ this is exactly the return amplitude for the quantum particle to be at the leftmost vertex of the graph as a function of time. In the next section, we will see how to use the solution to these equations to write a formula for the wave function on other vertices. For now, we will make a side comment. One way to solve these SD equations is to start with the free answer $\Sigma = 0$ and simply iterate the equations. The function $G_R(t)$ that we get after a finite number of iterations sums a set of diagrams where we cut off the recursive structure of the melon diagrams at some level. For example, after iterating zero times, we simply take the free propagator. After one iteration or two iterations, respectively, we are effectively summing diagrams of the form

 (3.26)

where all lines represent free propagators. Summing these diagrams is equivalent to evaluating the return amplitude for a particular cutoff version of the graph, where we keep all vertices that correspond to fan diagrams with

'recursive depth' equal to or less than the number of iterations of the SD equations. For example, the cutoffs corresponding to the zero-th, first and second iterations of the SD equations are shown in Fig. 3.2 in panels (b), (c) and (d).

This gives a perspective on why we get exponential decay of the two point function $\frac{2}{2^{N/2}}\mathrm{Tr}\,(\psi_1(t)\psi_1(0))$. For example, consider the self-energy diagrams on the left in (3.26). These describe oscillation between the operator ψ_1 and $J_{1abc}\psi_a\psi_b\psi_c$. In the graph picture, it represents a particle that is moving between the two states of the simple graph shown in (c) of Fig. 3.2. The result is a return amplitude that oscillates in time, $\cos\left(2^{1-\frac{q}{2}}Jt\right)$. If we consider the SD equations after two iterations, we are studying a particle moving in the somewhat more complicated graph shown in (d). It still oscillates, but the return amplitude has a somewhat lower average value. For any finite cutoff, or any finite iteration of the SD equations, we will get a correlator that oscillates in time. But in the limit where we study the infinite graph, the return amplitude decays exponentially because the particle can continue moving to the right forever in the infinite graph

2. Another important qualitative feature is that the expected size of the operator grows exponentially in time. This is because the degree of the graph is growing linearly with the generation k. The timescale for evolution from generation k to $k+1$ is proportional to the inverse of the degree, which is proportional to $1/k$. So as the particle moves farther out into the graph, it speeds up proportionally to its distance. This leads to the expectation value

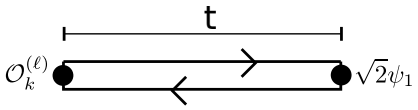
of k growing exponentially with time.[1]

3.4 Computing the wave function on the graph

In principle, one could evaluate the wave function for the particle moving on the graph by directly studying that problem. However, it is more convenient to translate the problem into a correlation function in the infinite temperature SYK model and then re-sum the 'melonic' SYK perturbation theory in the usual way.

Let's imagine that we want to compute the wave function that corresponds to the time evolution of the operator $\mathcal{O}_0(t) = 2^{\frac{1}{2}}\psi_1(t)$. We can write this explicitly as

$$\langle \mathcal{O}_k^{(\ell)} | e^{-i\widehat{H}t} | \mathcal{O}_0 \rangle = \frac{2^{\frac{1}{2}}}{2^{N/2}} \mathrm{Tr}\left(\mathcal{O}_k^{(\ell)} \ \psi_1(t) \right) = \tag{3.27}$$



The diagram in the last expression is the time contour for a path integral that evaluates the correlator. The two horizontal lines with arrows on them represent the forwards and backwards time evolution operators in the expression $\psi_1(t) = e^{iHt}\psi_1 e^{-iHt}$. In order to evaluate this quantity by perturbation theory, we should integrate interaction vertices of the SYK model everywhere on this folded time contour, connecting the loose propagators either to the ψ_1 operator at the right end, or the fermions in whatever $\mathcal{O}_k^{(\ell)}$ basis operator we are considering.

The simplest case is the return amplitude, when we take $\mathcal{O}_k^{(\ell)} = \mathcal{O}_0 \equiv 2^{\frac{1}{2}}\psi_1$, which we also represent with the symbol —. Then the quantity we are computing

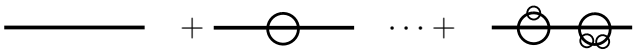
[1] To make this argument more reliably, one needs to know that the number of vertices at generation k is growing only exponentially in k, and not faster. The precise formula for the number of vertices at generation k is $\frac{1}{k}\binom{k(q-1)}{k-1}$, which grows exponentially in k for large k.

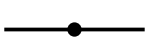
is

$$\langle \text{—} | e^{-i\widehat{H}t} | \text{—} \rangle = \frac{(2^{\frac{1}{2}})^2}{2^{N/2}} \mathrm{Tr}\left(\psi_1(0)\psi_1(t)\right) = 2G(t), \tag{3.28}$$

namely twice the two point function at infinite temperature for time separation t. If we like, we can write this (for $t > 0$) as the retarded propagator, since at infinite temperature $G_R(t) = \frac{1}{2^{N/2}} \mathrm{Tr}\left(\{\psi(t), \psi(0)\}\right) \theta(t) = 2\theta(t)G(t)$. So the answer for the return amplitude is given by the solution to the Schwinger-Dyson equations (3.25).

It is helpful to have a quick look at the perturbation theory that generates the SD equations. At large N, the perturbation theory for the return amplitude looks like the following

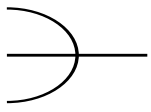
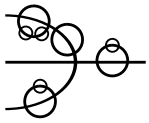
$\langle \text{—} | e^{-i\widehat{H}t} | \text{—} \rangle =$  (3.29)

$=$  (3.30)

Let us explain this notation. In the first line, in the Feynman diagrams, the two endpoints represent the operators $\sqrt{2}\psi_1$ inserted at time zero and time t. The lines in the Feynman diagrams represent free propagators, so the first diagram is simply $1 = (\sqrt{2})^2 \cdot \frac{1}{2}$, where the two factors of $\sqrt{2}$ are for the normalizations of the external operators, and the $\frac{1}{2}$ is a free fermion propagator. When we have interaction diagrams, we need to take care to sum over whether the interaction vertex is on the "forwards" or "backwards" portion of the time contour. It is easy to check that these contributions cancel unless all vertices are ordered in time in the same way that they are ordered in the diagram. In this case, the contributions from the two portions of the contour add together, giving an extra factor of two

for each vertex. So for example the second diagram gives $(\sqrt{2})^2 \cdot (2iJ)^2 \cdot \frac{t^2}{2} \cdot (\frac{1}{2})^5$. Here the $(2iJ)^2$ is for the two interaction vertices, the $\frac{t^2}{2}$ is for the integral over two ordered points between zero and t, and the $(\frac{1}{2})^5$ is for the five free fermion propagators. This evaluates to $-\frac{J^2t^2}{8}$. In the second line, we represent a dressed retarded propagator, which is equal to the return amplitude, as a line with a black dot in the middle.

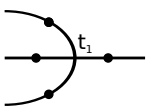
The next simplest case is when we take $\mathcal{O}_k^{(\ell)} = \mathcal{O}_1 =$. Now we need to evaluate a correlation function of a composite operator built out of three fermions, and the single-fermion operator $\psi_1(t)$. The lowest order diagram for this involves expanding down a single copy of the interaction vertex $J_{1abc}\psi_1\psi_a\psi_b\psi_c$, where the ψ_1 is contracted with our operator $\psi_1(t)$, and the other fermions are contracted with the $\mathcal{O}_1$ operator at time zero. Note that this Feynman diagram has the same structure as the fan diagram that we used to label the operator itself. At infinite N, the only other diagrams that contribute are 'melonic' decorations of this diagram, as in

$\langle$$|e^{-i\widehat{H}t}|$$\rangle =$  $+ \cdots +$  $+ \ldots$ (3.31)

These decorations can be summed by replacing the free propagators by dressed propagators that solve the Schwinger-Dyson equations. The full answer, including

the numerical factor from the normalization of the operators, is

$$\langle \text{  } |e^{-i\widehat{H}t}| \text{  } \rangle = -2^{1-\frac{q}{2}} iJ \int_0^t dt_1 G_R(t_1)^{q-1} G_R(t-t_1), \tag{3.32}$$

$$\langle \text{} |e^{-i\widehat{H}t}| \text{  } \rangle = -2^{1-\frac{q}{2}} iJ \int_0^t dt_1 \text{  } \tag{3.33}$$

This simple pattern persists for arbitrary operators $\mathcal{O}_k^{(\ell)}$: to compute the wave function, we can simply interpret the fan diagram of the operator $\mathcal{O}_k^{(\ell)}$ itself as a Feynman diagram, where all of the edges are dressed retarded propagators $G_R(t)$. We then integrate over the times of the vertices (subject to the ordering constraint which is imposed by the $\theta(t)$ in the retarded propagator). Including the correct numerical prefactor, one has for $k \geq 1$

$$\langle \mathcal{O}_k^{(\ell)} | e^{-i\widehat{H}t} | \text{  } \rangle = (-2^{1-\frac{q}{2}} iJ)^k \times \int dt_1 \dots dt_k \Big[G_R \text{ factors reflecting fan diagram} \Big]. \tag{3.34}$$

This gives an algorithm for computing the wave function of the particle moving on the graph, i.e. the time evolving operator. However there are two problems. First, in general we do not have an exact expression for the infinite temperature $G_R(t)$. Second, the number of different fan diagrams grows rapidly with generation k. In the special case of large q, both problems go away, because there is a known formula for $G_R(t)$, and as we will see, the different fan diagrams at a given generation are all proportional to the same function of time.

3.5 The wave function in the large-q SYK model

In this section we will evaluate the wave function and corresponding probability distribution $P_s(t)$ to leading nontrivial order in the large-q SYK model, namely $\frac{1}{q}$. To begin we will do a straightforward large q analysis, where t does not scale with q. This approximation breaks down at times of order q, and we will comment on how to resum t/q effects at the end of the section.

At large q and infinite temperature, there is a simple expression for the product of q propagators [2]

$$G_R(t)^q = \frac{\theta(t)}{\cosh^2 \mathcal{J}t} + O(1/q). \tag{3.35}$$

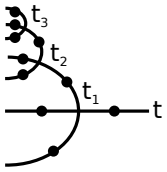
Taking the $1/q$-th power of this, we find that

$$G_R(t) = \theta(t) + O(1/q), \tag{3.36}$$

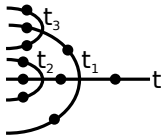
so a single propagator is almost given by the free answer. The fact that $G_R(t)^q$ is nontrivial will lead to an interesting wave function. However, the computation will be simplified by the fact that any fixed $O(1)$ number of propagators are simply step functions, which means that once the ordering of time arguments are imposed we can set them equal to unity. A very useful point is that this implies that the wave function has the same time dependence for all operators of a given generation k. This means we only have to compute a single representative r for each k. This will make the computation of the wave function tractable.

Let's understand how this works by considering the expressions for two differ-

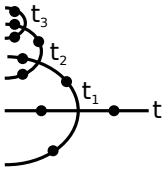
ent generation 3 operators, each "born" from $|$  $\rangle$. The first expands "depth-first,"

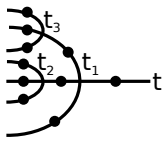
$$\Big\langle \text{  } \Big| e^{-i\widehat{H}t} \Big| \text{  } \Big\rangle \propto \int dt_1\, dt_2\, dt_3 \quad \text{  } \qquad (0 < t_3 < t_2 < t_1 < t), \tag{3.37}$$

and the second expands "breadth-first,"

$$\Big\langle \text{  } \Big| e^{-i\widehat{H}t} \Big| \text{  } \Big\rangle \propto \int dt_1\, dt_2\, dt_3 \quad \text{  } \qquad (0 < t_3, t_2 < t_1 < t). \tag{3.38}$$

We have not included constants of proportionality, because we will fix them below by a different argument. The propagators with a dot on them represent the dressed retarded propagators G_R, so that explicitly


$$= G_R(t_3)^{q-1} G_R(t_3 - t_2) G_R(t_2)^{q-2} G_R(t_2 - t_1) G_R(t_1)^{q-2} G_R(t_1 - t), \tag{3.39}$$


$$= G_R(t_3)^{q-1} G_R(t_2)^{q-1} G_R(t_3 - t_1) G_R(t_2 - t_1) G_R(t_1)^{q-3} G_R(t_1 - t). \tag{3.40}$$

The ratio of these expressions is $G_R(t_3 - t_2)G_R(t_1)/G_R(t_2)G_R(t_3 - t_1)$. Since $G_R(t) = \theta(t)$ at leading order in large q, we see that the nontrivial time dependence of these integrands is equal. The only difference is that they have a different

set of θ functions that impose different orderings of the time arguments. This applies more generally: we can write the integrand of any dressed fan diagram at generation k in the simple form

$$\prod_{j=1}^{k} \frac{1}{\cosh^2 \mathcal{J} t_j} + O(1/q), \tag{3.41}$$

times a set of step functions that impose the ordering of time arguments appropriate for a given fan diagram.

We now have to do the integral. In principle, this integral should be over times $t_1 \dots t_k$ respecting the constraints from the step functions. In our example, the depth-first operator has $t_3 < t_2 < t_1$, and the breadth-first operator has $t_3 < t_1$ and $t_2 < t_1$, with no relationship between t_2 and t_3. However, since the integrand (3.41) is symmetric under interchanges of the t_j, these restrictions will only affect the numerical prefactor and not the time dependence of the result. Thus, the time dependence of any dressed fan diagram will be the same. Picking the "depth-first" expansion to be our representative r at each generation k, we have that

$$\langle \mathcal{O}_k^{(r)} | e^{-i\widehat{H}t} | \text{—} \rangle \propto \left[\int_0^t \frac{dt_j}{\cosh^2 \mathcal{J} t_j} \right]^k \propto \tanh^k \mathcal{J} t. \tag{3.42}$$

This implies that at leading order in large q, we have $P_{s_k}(t) \propto \tanh^{2k}(\mathcal{J}t)$, where $s_k = 1 + (q-2)k$.

As a final step, we need to determine the numerical coefficients. We can do this by requiring that the probability distribution remain normalized for all times. The trick here is to use the fact that we already have an expression for $P_1(t) = G_R(t)^2$

from (3.35) that is accurate at first subleading order in the $1/q$ expansion:

$$P_1(t) = 1 - \frac{4}{q} \log\cosh \mathcal{J}t + O(1/q^2). \tag{3.43}$$

Now, to determine the numerical coefficients for P_{s_k} with $k > 1$, we try to solve

$$1 = P_1(t) + \sum_{k=1}^{\infty} \mathcal{N}_k \tanh^{2k} \mathcal{J}t, \tag{3.44}$$

to order $1/q$. Indeed, one can solve this equation by setting $\mathcal{N}_k = 2/kq$. This gives the probability distribution at leading nontrivial order in $1/q$

$$P_s(t) = \begin{cases} 1 - \frac{4}{q} \log\cosh \mathcal{J}t + O(1/q^2), & s = 1, \\ \frac{2}{kq} \tanh^{2k} \mathcal{J}t + O(1/q^2), & s = 1 + (q-2)k, \quad k = 1, 2, 3, \ldots . \end{cases} \tag{3.45}$$

We will now make several comments about this result.

1. One can evaluate the expectation value of the size s in this distribution. At leading order in $1/q$, we have:

$$\langle s \rangle = \sum_k P_{s_k} s_k = 1 + \sum_{k=1}^{\infty} 2 \tanh^{2k} \mathcal{J}t \tag{3.46}$$

$$= \cosh(2\mathcal{J}t). \tag{3.47}$$

   This result for the expected value of the size determines the initial exponential growth of the anticommutator-squared, via (3.12). We conclude that the chaos exponent at large q and infinite temperature is $\lambda_L = 2\mathcal{J}$. The formulas from [2] can be used to show that in the large-q model we have

$\lambda_L = 2\mathcal{J}\sqrt{1-x^2}$ where $x = \frac{q^2 E}{\mathcal{J}N}$. Here the energy spectrum is such that $-1 < x < 1$, and $x = 0$ corresponds to infinite temperature state. So we find agreement with previous results.

2. Note that the leading-order answer for (3.46) depends on the $\frac{1}{q}$-suppressed probabilities for $s > 1$, because $s \approx qk$ and this factor of q cancels against the $\frac{1}{q}$ suppression. In other words, at large q, the operator initially has only small probability (of order $\frac{1}{q}$) to grow, but if it does grow it gets so big (size proportional to q) that this makes a large effect on the expected value of the size. This is reflected in the fractional variance of the size distribution, which is large, proportional to q

$$\frac{\langle s^2\rangle - \langle s\rangle^2}{\langle s\rangle^2} = \frac{q}{2}\tanh^2 2\mathcal{J}t + O(1). \tag{3.48}$$

3. As another example of something one can compute with this distribution, we can generalize the logic that led to (3.12) slightly, finding

$$\frac{1}{N}\sum_j \frac{1}{2^{N/2}}\mathrm{Tr}\left[\{\psi_1(t),\psi_j\}\{\psi_1(t'),\psi_j\}\right] = \sum_s \frac{s}{N}\sqrt{P_s(t)P_s(t')}. \tag{3.49}$$

Evaluating this with our large q result (3.45), we find

$$\frac{1}{N}\sum_j \frac{1}{2^{N/2}}\mathrm{Tr}\left[\{\psi_1(t),\psi_j\}\{\psi_1(t'),\psi_j\}\right] = \frac{1}{N}\frac{\cosh\left[\mathcal{J}(t+t')\right]}{\cosh\left[\mathcal{J}(t-t')\right]} + \ldots \tag{3.50}$$

4. So far, we have considered a simple large-q limit, where we do not allow t to scale with q. This approximation will break down at times of order

q. It would be nice to extend our analysis to resum effects of order t/q. Although we have not studied this systematically, we will make a few comments. To capture the important effects, one can no longer approximate $G_R(t)$ as simply $\theta(t)$ in cases where the time argument can be long. Instead, we can approximate it as $G_R(t) = \theta(t)\cosh^{-2/q}(\mathcal{J}t) \approx \theta(t)e^{-2\mathcal{J}t/q}$. We expect based on numerics and [101] that this expression is accurate for all time t, although it does not follow from the approximation worked out in [2].

A convenient feature of this approximation is that (ignoring the step functions) we have $G_R(t)G_R(t') = G_R(t+t')$. This composition property allows us to convert fan diagram integrands into each other, so we retain the property that only one representative from each generation must be computed. In the example given above, to convert (3.39) to (3.40), we use $G_R(t_3-t_2)G_R(t_1) = G_R(t_3-t_1)G_R(t_2)$. Another simplification is that inside the integrand, we can expect to approximate $G_R(t_j)^{q-\alpha} \approx G_R(t_j)^q$, because the presence of the factor $G_R(t_j)^q$ will make the integral prefer the region where t_j is of order one, so the factor $G_R(t_j)^\alpha$ will be approximately one.

Following these approximations, we find that the total effect is to multiply the wave function (3.42) by $G_R(t)$. Normalizing the probability distribution, we find the expression

$$P_s(t) = \frac{\Gamma(k+2/q)}{\Gamma(k+1)\Gamma(2/q)}\frac{\tanh^{2k}\mathcal{J}t}{\cosh^{4/q}\mathcal{J}t}, \qquad s = 1+(q-2)k. \tag{3.51}$$

This probability distribution resums the t/q corrections to our straightforward large-q result (3.45). However, it is not fully satisfactory because there

are expected to be k/q corrections that are not accurately summed here. We hope that the expression is nevertheless qualitatively accurate even for large k. Note that at our level of approximation the denominator could be written $e^{4\mathcal{J}t/q}$ instead of $\cosh^{4/q}\mathcal{J}t$.

Our main purpose in writing the expression (3.51) is that one finds a very similar formula in a classical model of operator growth discussed in § 3.2.1.

5. It is sometimes convenient to define a "coarse-grained" wave function by $\Psi_s(t) = (-i)^k\sqrt{P_s(t)}$. This is the amplitude for $\psi_1(t)$ to be of size s at time t. We ought to have a composition property where the two point function of fermions can be computed by inserting a complete set of states at an intermediate time and summing over the sizes of all operators that appear. In other words, we should have

$$\Psi_1(t_1+t_2) = \langle\psi_1(t_1+t_2)|\psi_1(0)\rangle = \sum_s \langle\psi_1(t_1+t_2)|\mathcal{O}_s(t_2)\rangle\langle\mathcal{O}_s(t_2)|\psi_1(0)\rangle,$$
$$= \sum_s \Psi_s^*(t_1)\Psi_s(-t_2). \tag{3.52}$$

Indeed, one can check that this property holds for (3.51). It follows that it also holds at order $\frac{1}{q}$ for (3.45).

3.6 Discussion

In this paper, we discussed the time evolution of a simple fundamental fermion operator ψ_1 in the SYK model. In the large N limit, we related the operator growth problem to the problem of a particle moving on a rapidly expanding graph.

We computed the size distribution for the evolving operator explicitly in two cases: numerically for $N = 30$ fermions with $q = 4$ and analytically at large N and large q. We showed how to use this size distribution to compute out-of-time-order correlators at infinite temperature.

Throughout, we have emphasized a particular decomposition of the time evolving operator, into components with a given size (number of elementary fermions appearing in a product). We would like to contrast this with the Baker-Campbell-Hausdorff expansion

$$\psi_1(t) = e^{iHt}\psi_1 e^{-iHt} = \psi_1 + it[H,\psi_1] - \frac{t^2}{2}[H,[H,\psi_1]] - \frac{it^3}{3!}[H,[H,[H,\psi_1]]] + \dots . \tag{3.53}$$

These terms also give a decomposition of the time-evolving operator. We emphasize that this is different from the size decomposition, because the terms at order k in the BCH expansion do not all have the same size. While the kth nested commutator contains terms up to size $s = k(q-2)+1$, there is also weight on operators of shorter sizes. For instance, in the $q = 4$ SYK model, the $k = 2$ term in the BCH expansion contains operators of size 5 as well as an operator of size 1, namely ψ_1. The fact that our wave function $\sqrt{P_s(t)}$ is a nontrivial function of time indicates that it receives contributions from many different orders in the BCH expansion (starting at order k, where $s = 1 + (q-2)k$).

Another point we would like to emphasize is the following. The notion of size that we have used makes explicit reference to a particular set of simple operators $\{\psi_i\}$, out of which we construct complicated ones. This set of simple operators is determined by the Hamiltonian itself, and it depends on the q-local and sparse

nature of H. If instead the Hamiltonian were a totally random matrix, we would have no sensible notion of simple operators, and no good way to define size. However, for Hamiltonians such as SYK, a preferred set of simple operators is selected by the fact that the interaction can be written in terms of finite (order q) products of them.

There are many possible directions for improvements on our work. For example, it would be interesting to understand $1/N$ corrections to the $P_s(t)$ distribution at a level that would make it possible to see saturation of the late-time distribution. Another challenge is to extend the approach studied here to compute out-of-time-order correlators at finite temperature. For instance, one might be tempted to try to define a size distribution $P_s^{(\beta)}(t)$ with respect to an inner product $(A, B)_\beta \equiv Z(\beta)^{-1} \operatorname{Tr}[A^\dagger e^{-\beta H/2} B e^{-\beta H/2}]$, where $Z(\beta)$ is the thermal partition function. In principle, one could use the Schwinger-Dyson equations at large q to compute a candidate wave function. However, this necessarily requires the use of a different set of operators $\mathcal{O}_k^{(\ell)}(\beta)$ that now depend on the temperature β. Unfortunately these operators do not appear to admit a simple relationship between "generation" k and operator size s. This means we do not know how to extract the expected size from this candidate distribution, and we do not know how to relate it to the out-of-time-order commutator as we did in § 3.2.3.

In holographic theories, operator growth is described by a particle falling towards the black hole horizon. It is tempting to think of the radial direction in the graph of operators as being similar to the radial direction in the bulk theory, so that the particle propagating deeper into the graph resembles the particle falling into the bulk. We do not know if there is a more precise connection to be made

there.

Chapter 4

Finite Temperature Operator Growth

In many-body chaotic systems, the size of an operator generically grows in Heisenberg evolution, which can be measured by certain out-of-time-ordered four-point functions. However, these only provide a coarse probe of the full underlying operator growth structure. In this article we develop a methodology to derive the full growth structure of fermionic systems, that also naturally introduces the effect of finite temperature. We then apply our methodology to the SYK model, which features all-to-all q-body interactions. We derive the full operator growth structure in the large q limit at all temperatures. We see that its temperature dependence has a remarkably simple form consistent with the slowing down of scrambling as temperature is decreased. Furthermore, our finite-temperature scrambling results can be modeled by a modified epidemic model, where the thermal state serves as a vaccinated population, thereby slowing the overall rate of infection.

4.1 Introduction & Summary

In chaotic quantum many-body systems, operators grow in size as time evolves. For example, in spatially local systems one expects that the extent of an operator $\mathcal{O}(t)$ grows as

$$\frac{d}{dt}\text{Volume}\left[\mathcal{O}(t)\right] \propto \text{Surface Area}\left[\mathcal{O}(t)\right] \tag{4.1}$$

since the new terms generated by taking $[H, \mathcal{O}(t)]$ will live on the boundary of the domain of $\mathcal{O}(t)$ [80–82, 84]. Consequently, the extent grows linearly with an effective "speed of light" $\simeq vt$. Up to exponential error all operators outside the effective light-cone will commute with $\mathcal{O}(t)$. This effective speed of light is known as the Lieb-Robinson velocity [80]. This highlights the fact that space can be a derived concept in quantum mechanics, as without the Hamiltonian there may not be a sense in which one piece of the Hilbert space factorization is closer to another.

Now we would like to contrast this behavior with that exhibited by q-local systems, where the Hamiltonian couples all the degrees of freedom together in q-body interactions. Consequently, there is no notion of spatial locality, and we accordingly refer to such interactions as coupling together "internal" degrees of freedom. Yet, there remains structure in the evolution of operators in these systems, as we often find that the sizes of operators grow exponentially

$$\frac{d}{dt}\text{Size}\left[\mathcal{O}(t)\right] \propto \text{Size}\left[\mathcal{O}(t)\right] \tag{4.2}$$

where by size we mean the number of simple operators multiplied together in a

typical piece of $\mathcal{O}(t)$. The intuition behind this growth is that the percentage of the q-body interactions utilized in $[H, \mathcal{O}(t)]$ is proportional to the size of $\mathcal{O}(t)$, and almost all the resultant operators obtained from $[H, \mathcal{O}(t)]$ are bigger than $\mathcal{O}(t)$ [7, 63, 91–93, 102, 103].[1]

Systems with both spatial locality and a large number of internal degrees of freedom–such as (chaotic) field theories in the large-N limit – display both linear spatial growth and exponential internal size growth [23, 89, 90, 104, 105]. The growth of an evolving simple operator $W(t)$ can be probed using another simple operator V, using the (anti-)commutator squared $\langle |[W(t), V]|^2 \rangle$ or their corresponding out-of-time-order correlator (OTOC) $\langle W^\dagger(t) V^\dagger W(t) V \rangle$ [21, 23, 27, 83, 86, 94, 96, 97, 106, 107].

In order to develop the coarse-grained profile of operator growth, one must compute many OTOCs. The "chaos bound" [5] obeyed by OTOCs suggests that after an initial dissipation time, they de-correlate no faster than exponentially, with a rate λ_L no larger than $2\pi T$ where $T = 1/\beta$ is the temperature. This implies that presence of the thermal state $\rho \propto \exp(-\beta H)$ slows down the effective growth rate of operators as temperature is decreased.

The Heisenberg evolution of operator $\mathcal{O}(t)$ is independent of temperature, so the entire effect of temperature must be contained in the matrix elements of $\mathcal{O}$. Therefore, the natural finite temperature generalization of operator size has remained an open question (one recent proposal is given in [108]).

In this article, we address this issue by characterizing not only the average size of an operator but its entire size distribution. We then can define the effective

[1] Eq. (4.2) actually can be considered as the same formula as in Eq. (4.1) applied to a completely connected graph, so that the area of a region is proportional to its volume [102].

size distribution of an operator at finite temperature by how it *changes* the size of the square root of thermal density operator $\rho^{1/2}$ (we explain why this is a natural choice in section (4.3)).

This definition leads to some nontrivial general results, independent of the details of the specific physical system. In particular, we observe that in generic fermion systems, the effective size of a single fermion operator is "thermally renormalized" to a value $\delta_\beta = G(\beta/2)$ smaller than 1, where $G(\tau)$ is the thermal two-point function. The size of the thermal operator $\rho^{1/2}$ itself is $\frac{N}{2}(1-\delta_\beta)$, determined by the same renormalization factor δ_β. To gain a more explicit understanding, we will work in the context of the SYK model [4,98], a q-local Hamiltonian built out of N flavors of Majorana fermions, which saturates the chaos bound at low temperatures [2,14,97]

The remainder of the article is organized as follows. We begin in section (4.2) by building up the notion of operator "size". First, we show that one may expand any operator $\mathcal{O}(t)$ in an orthonormal operator basis of the unique products of Majorana flavors. In the doubled theory, the operator basis maps to an orthonormal basis of states in the doubled Hilbert space. We define a "size" operator n in the doubled theory counting the average number of flavors in an operator basis state. We are then able to demonstrate that four-point functions measure the average "size" of an operator. Therefore, the de-correlation of a thermal OTOC is exactly equivalent to the growing average size of the operator $\psi_1(t)\rho^{1/2}$

$$-\frac{1}{N}\sum_{j=1}^{N}\mathrm{Tr}\left(\rho^{1/2}\psi_1(t)\psi_j\psi_1(t)\rho^{1/2}\psi_j\right) = 1-\frac{2}{N}n\left[\psi_1(t)\rho^{1/2}\right]$$

The average size of the operator $\psi_1(t)\rho^{1/2}$ starts at $n\left[\psi_1\rho^{1/2}\right] \approx n\left[\rho^{1/2}\right] = \frac{N}{2}(1-\delta_\beta)$ and then grows sigmoidally in time, eventually saturating (scrambling) at a value of $N/2$.

Up to this point we have been discussing the average size of the operator $\psi(t)\rho^{1/2}$. In fact, the entire size distribution of this operator has physical significance. Hence, in section (4.3) we construct generating functions for operator size distributions by inserting a weighting factor $\exp(-\mu n)$. First, we study the size distribution of the thermal operator $\rho^{1/2}$ by setting up a generating function $\mathcal{Z}_\mu\left[\rho^{1/2}\right]$, which is similar to a grand canonical partition function. Next, we show that the fractional distance to scrambling for the operator $\rho^{1/2}$ is always given by $\delta_\beta \equiv 1 - n/n_* = G(\beta/2) < 1$. Then, we set up the generating function for the size distribution of $\psi(t)\rho^{1/2}$, which we find naturally splits into a product of $\mathcal{Z}_\mu\left[\rho^{1/2}\right]$ and a modified two-point function $\mathcal{G}_\mu(t)$. We show that the μ-expansion of $\mathcal{G}_\mu(t)$ determines the growth distribution induced by multiplying $\rho^{1/2}$ by $\psi(t)$, and that $\mathcal{G}_\mu(t)$ is simply the two-point function for the original theory with a μ-dependent twisted boundary condition. We conclude the section by noting that on average, the size increase induced by a single fermion is given by the fractional scrambling distance δ_β, which leads us to propose that δ_β should be interpreted as a thermally renormalized unit of size.

Everything in sections (4.2) and (4.3) applies to general fermionic systems. In sections (4.4) and (4.5), we apply this methodology to the large-q SYK model. Solving the large-N saddle point equation in the large q limit with our μ-dependent twist, we obtain the full operator growth structure. After a dynamical renormalization of coupling constant $\mathcal{J}$ (which after a short amount of time essentially

amounts to replacing the coupling with a smaller β-dependent constant) and a renormalization of size unit from 1 to δ_β, we observe that the full growth structure has the same functional form as the infinite temperature case. The dynamical renormalization of the coupling is the signature of the slowdown of the effective growth rate as temperature is decreased.

We conclude the section by discussing how to understand this finite temperature slowdown of scrambling in an epidemic model, where the thermal factor effectively vaccinates a large subset of the population, thereby slowing down the overall infection rate. We end the paper by discussing implications and future directions in section (4.6).

4.2 Operator Distributions and Two-Sided Wavefunctions

As an operator $\mathcal{O}(t)$ evolves in time, it becomes supported along operators of increasing size. This can be inferred from the Heisenberg equation of motion $\dot{\mathcal{O}}(t) = i[H, \mathcal{O}(t)]$. Now, in order to properly discuss how much one operator is supported along another, we need an operator inner product. When the Hilbert space is finite-dimensional it is natural to use the Frobenius inner product: $\langle \mathcal{O}_A | \mathcal{O}_B \rangle \equiv \text{Tr}(\mathcal{O}_A^\dagger \mathcal{O}_B)$. We may then expand operators in an orthonormal operator basis, which amounts to inserting a complete set of operators $\{\Gamma_I\}$

$$\mathcal{O}(t) = \sum_I \Gamma_I \text{Tr}\left(\Gamma_I^\dagger \mathcal{O}(t)\right) \equiv \sum_I c_I(t)\, \Gamma_I$$

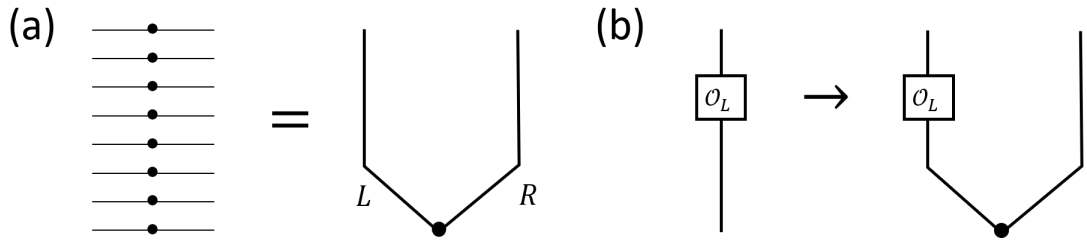


Figure 4.1: Illustration of the purification procedure that maps operators to states in a doubled Hilbert space. (a) A maximally entangled state $|0\rangle$ (Eq. (4.3) which can be viewed as many EPR pairs between the two systems. (b) The mapping between operator $\mathcal{O}$ and the corresponding state $|\mathcal{O}\rangle$ obtained by applying $\mathcal{O}$ to the left system (see Eq. (4.4)).

Note that at this point we have set up a Hilbert space of operators. If the original Hilbert space $\mathbb{H}$ has dimension L, the operator Hilbert space is $\mathbb{H} \otimes \overline{\mathbb{H}}$, with dimension L^2.

4.2.1 Purification

Since $\mathbb{H} \otimes \overline{\mathbb{H}}$ is isomorphic to $\mathbb{H} \otimes \mathbb{H}$, one can always maps each operator to a quantum state in a "doubled" system with Hilbert space dimension L^2. More explicitly, this mapping is defined by considering two copies of the original physical system, named as L and R, and introducing a maximally entangled state $|0\rangle$ (see Fig. 4.1(a)). For any maximally entangled state, there is a basis choice of the form $\{|n\rangle_L \otimes |m\rangle_R\}$ such that

$$|0\rangle = \sum_{m,n} \delta_{mn} |n\rangle_L \otimes |m\rangle_R = \sum_n |n\rangle_L \otimes |n\rangle_R \tag{4.3}$$

For later convenience we have chosen the norm of the state to be $\langle 0|0\rangle = L$.

Then the operator-to-state mapping is defined by

$$\mathcal{O} \to |\mathcal{O}\rangle \equiv \mathcal{O}_L \otimes \mathbb{I}_R |0\rangle \tag{4.4}$$

where $\mathcal{O}_L$ is the operator $\mathcal{O}$ acting on the Hilbert space of the left system, as is illustrated in Fig. 4.1(b). It is easy to verify that the inner product ${}_{LR}\langle \mathcal{O}_A | \mathcal{O}_B \rangle_{LR} = \mathrm{Tr}(\mathcal{O}_A^\dagger \mathcal{O}_B)$ is determined by the Frobenius inner product of the corresponding operators. Our orthonormal basis of operators $\{\Gamma_I\}$ will thereby serve as an orthonormal basis of states $|\Gamma_I\rangle$. Thus, the problem of understanding how $\mathcal{O}(t)$ is distributed across a particular choice of basis operators is equivalent to understanding how the two-sided state $|\mathcal{O}(t)\rangle$ is distributed across a particular choice of two-sided basis states. Since the choice of maximally entangled state $|0\rangle$ is not unique, the operator-to-state mapping has an ambiguity of $U(L)$ acting on the R system. Since the same transformation is performed to $|\mathcal{O}\rangle$ and the basis vector $|\Gamma_I\rangle$, all our discussion will be independent from this freedom of basis choice.

4.2.1.1 Orthonormal Basis of Operators

One of the simplest algebras with an interesting finite dimensional representation is the algebra of N flavors of Majorana fermions, where N is even:

$$\{\psi_i, \psi_j\} = 2\delta_{ij} \tag{4.5}$$

Note that this implies that $\psi_j^2 = 1$, which will be convenient for our purposes, unlike the more common convention where $\{\psi_i, \psi_j\} = \delta_{ij}$ and thus $\psi_j = 1/2$.

Such operators are traceless, Hermitian, and unitary. Furthermore, the algebra is invariant under taking any single $\psi_i \to -\psi_i$, and so the product of any subset of the N fermions is also traceless. Thus, it is easy to construct an orthogonal operator basis by taking unique ordered products of Majorana fermions

$$\Gamma_I \equiv \Gamma_{i_1 i_2 ... i_k} = i^{\frac{k(k-1)}{2}} \psi_{i_1} ... \psi_{i_k} \qquad 1 \leq i_1 < i_2 < ... < i_k \leq N \tag{4.6}$$

where the pre-factor has been inserted so that the resultant Γ_I matrices are Hermitian. All nontrivial Γ_I (with $k > 0$) are traceless. Since the product $\Gamma_I \Gamma_J$ is also a string of fermions, which is only trivial when $I = J$ (when the two strings are identical and Majorana fermions pairwise cancel), we have

$$\mathrm{Tr}\left(\Gamma_I^\dagger \Gamma_J\right) = \mathrm{Tr}\left(\Gamma_I \Gamma_J\right) = \delta_{IJ} \mathrm{Tr}\left(1\right) \tag{4.7}$$

Furthermore, the basis operators Γ_I have simple algebraic relations, since they either commute or anti-commute according to the relation

$$\Gamma_I \Gamma_J = (-1)^{|I||J| + |I \cap J|} \Gamma_J \Gamma_I$$

where $|I|$ is the number of elements in the multi-index I.

4.2.1.2 Mapping Basis Operators to Basis States

The purification isomorphism is quite simple to realize, as has been discussed by [109, 110]. We consider two copies of the original system, which contains $2N$ Majorana fermions labeled by ψ_j^L and ψ_j^R, $j = 1, 2, ..., N$. We then define a

maximally entangled state $|0\rangle$,

$$\left(\psi_j^L + i\psi_j^R\right)|0\rangle = 0, \ \forall j \tag{4.8}$$

We may think of this state as a vacuum (all spins down, all bits set to 0) with regards to a set of entangled complex fermions operators

$$c_j|0\rangle = 0 \qquad c_j \equiv \frac{\psi_j^L + i\psi_j^R}{2} \qquad \{c_j, c_k\} = \left\{c_j^\dagger, c_k^\dagger\right\} = 0 \qquad \left\{c_j, c_k^\dagger\right\} = \delta_{jk}$$

where $c_j^\dagger = (c_j)^\dagger$. Since state $|0\rangle$ is the ground state of a quadratic Hamiltonian $H = \sum_j c_j^\dagger c_j$, it is straightforward to compute the entanglement entropy [111] and verify that the state is maximally entangled between L and R. As we discussed earlier, the choice of $|0\rangle$ is not unique, but this choice is convenient for our purpose. The basis operators Γ_I are mapped to states in the doubled system of $2N$ Majorana fermions:

$$|\Gamma_I\rangle \equiv \Gamma_I^L|0\rangle = i^{\frac{k(k-1)}{2}}\psi_{i_1}^L ... \psi_{i_k}^L|0\rangle = i^{\frac{k(k-1)}{2}}c_{i_1}^\dagger ... c_{i_k}^\dagger|0\rangle = c_{i_k}^\dagger ... c_{i_1}^\dagger|0\rangle \tag{4.9}$$

Therefore each basis operator Γ_I is mapped to a particular fermion configuration in the doubled system, with fermions $i_1, i_2, ..., i_k$, as is illustrated in Fig. 4.2(a). Essentially, the identity operator maps to the vacuum and nontrivial operators are mapped to excitations in the doubled theory.

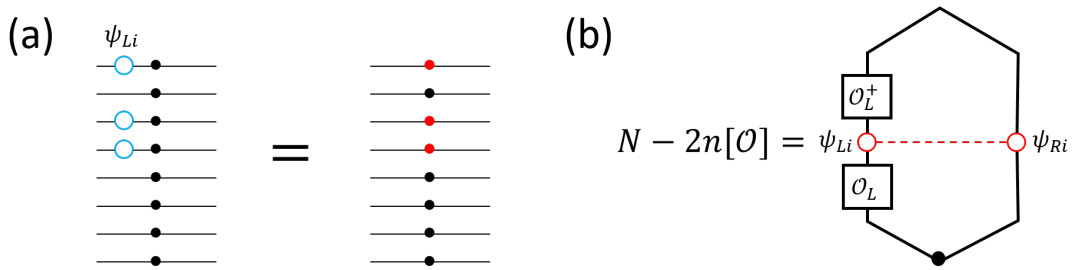


Figure 4.2: (a) The mapping of a Majorana string Γ_I in Eq. (4.6) to a state in the doubled system. Each fermion operator ψ_{Li} creates a fermion (red dot) while the fermions that are absent in Γ_I stays in the vacuum state with fermion number 0 (black dot). (b) Illustration of the relation between average size of operator $\mathcal{O}$ and OTOC.

4.2.2 Four-Point Functions Probe Operator Size

At this point, we can discuss the number operator $n_j \equiv c_j^\dagger c_j$, which returns 1 when applied to basis states containing the flavor j and zero otherwise

$$n_j \equiv c_j^\dagger c_j = \frac{1}{2}\left(1 + i\psi_j^L \psi_j^R\right) \qquad \langle \Gamma_I | n_j | \Gamma_J \rangle = \delta_{j\in J} \langle \Gamma_I | \Gamma_J \rangle \tag{4.10}$$

Thus, we see that for a generic operator $\mathcal{O}$, the expectation value of n_j returns the percentage of basis operators in $\mathcal{O}$ containing flavor j. Furthermore, we note that this expectation value is closely related to a one-sided four-point function (see Fig. 4.2), since

$$\begin{aligned} \langle \mathcal{O}|(2n_j - 1)|\mathcal{O}\rangle &= \langle \mathcal{O}| i\psi_j^L \psi_j^R |\mathcal{O}\rangle = -\langle 0|\left(\mathcal{O}^L\right)^\dagger \psi_j^L \mathcal{O}^L i\psi_j^R|0\rangle \\ &= \langle 0|\left(\mathcal{O}^L\right)^\dagger \psi_j^L \mathcal{O}^L \psi_j^L|0\rangle = \mathrm{Tr}_L\left(\left(\mathcal{O}^L\right)^\dagger \psi_j^L \mathcal{O}^L \psi_j^L\right) \\ \Rightarrow \langle \mathcal{O}|(2n_j - 1)|\mathcal{O}\rangle &= \mathrm{Tr}\left(\mathcal{O}^\dagger \psi_j \mathcal{O} \psi_j\right) \end{aligned} \tag{4.11}$$

Here we have assumed $\mathcal{O}$ to be fermionic. In the first two steps, we simply plugged in the definitions of n_j and $|\mathcal{O}\rangle$. In the third step, we anti-commuted $i\psi_j^R$ through $\mathcal{O}^L$, as right fermionic operators anti-commute with left fermionic operators. Then, we used the definition of $|0\rangle$ (4.8) to replace $-i\psi_j^R|0\rangle$ with $\psi_j^L|0\rangle$. Afterwards, we had an expectation value of only left operators for a maximally entangled state, so we traced out the right Hilbert space entirely, leaving us with an infinite temperature four-point function of the left-only system.

The relationship between operator quantities and one-sided correlators is simpler in terms of the anti-commutator squared, since we have

$$\begin{aligned} \frac{1}{4}\mathrm{Tr}\left(\{\mathcal{O},\psi_j\}^\dagger\{\mathcal{O},\psi_j\}\right) &= \frac{1}{2}\langle\mathcal{O}|\mathcal{O}\rangle + \frac{1}{2}\langle\mathcal{O}|(2n_j-1)|\mathcal{O}\rangle \\ \Rightarrow \frac{1}{4}\mathrm{Tr}\left(\{\mathcal{O},\psi_j\}^\dagger\{\mathcal{O},\psi_j\}\right) &= \langle\mathcal{O}|n_j|\mathcal{O}\rangle \equiv n_j[\mathcal{O}] \end{aligned} \tag{4.12}$$

where we used (4.11) to replace $\mathrm{Tr}\left(\mathcal{O}^\dagger\psi_j\mathcal{O}\psi_j\right)$ with $(2n_j-1)$. One should note that if $\mathcal{O}$ is bosonic, the right-hand side of (4.11) will acquire a minus sign and the anti-commutators in (4.12) will be replaced with commutators. We denote the average value of n_j in operator $\mathcal{O}$ as $n_j[\mathcal{O}]$.

We can also define a total number operator (a.k.a. size operator) that returns the number of flavors or size of a basis state

$$n \equiv \sum_{j=1}^{N} n_j = \sum_{j=1}^{N} c_j^\dagger c_j \qquad\qquad \langle\Gamma_I|n|\Gamma_J\rangle = |I|\langle\Gamma_I|\Gamma_J\rangle \tag{4.13}$$

with $|I|$ the number of Majorana fermion operators in the string Γ_I. Consequently, $\langle\mathcal{O}|n|\mathcal{O}\rangle$ is the average number of flavors in the operator $\mathcal{O}$, or the average size of

the operator $\mathcal{O}$. By flavor averaging Eq. (4.12), we see that the flavor-averaged anti-commutator squared measures the average size of the operator $\mathcal{O}$

$$\frac{1}{4N}\sum_{j=1}^{N}\mathrm{Tr}\left(\{\mathcal{O},\psi_j\}^\dagger\{\mathcal{O},\psi_j\}\right)=\frac{\langle\mathcal{O}|n|\mathcal{O}\rangle}{N}\equiv\frac{n\left[\mathcal{O}\right]}{N}$$

where the anti-commutators are replaced with commutators if $\mathcal{O}$ is bosonic.

Alternatively, we may flavor average Eq. (4.11) in order to relate the flavor-averaged four-point function to the average size

$$\frac{(-1)^{|\mathcal{O}|}}{N}\sum_{j=1}^{N}\mathrm{Tr}\left(\mathcal{O}^\dagger\psi_j\mathcal{O}\psi_j\right)=\langle\mathcal{O}|\left(1-\frac{2n}{N}\right)|\mathcal{O}\rangle\equiv\langle\mathcal{O}|\delta|\mathcal{O}\rangle\equiv\delta\left[\mathcal{O}\right]\tag{4.14}$$

where $(-1)^{|\mathcal{O}|}$ is 1 if $\mathcal{O}$ is bosonic and -1 if $\mathcal{O}$ is fermionic. Note that a totally scrambled operator (i.e. a random Hermitian operator) has a size $n_*=N/2$, so we see that the flavor-averaged four-point function measures the average fractional distance $1-2n/N$ an operator's size is from this scrambled value. Therefore we define the fractional scrambling distance operator

$$\delta\equiv1-\frac{n}{n_*}=1-\frac{2n}{N}\tag{4.15}$$

4.2.3 Operator Size Generating Function

By defining the number operator n, we can now go beyond the average operator size probed by four-point functions. Rather than just the average, we study all

moments systematically by introducing a generating function [103]

$$\mathcal{Z}_\mu[\mathcal{O}] = \langle \mathcal{O}|e^{-\mu n}|\mathcal{O}\rangle$$

By taking derivatives of the generating function we can obtain all moments of n:

$$\langle \mathcal{O}|n^k|\mathcal{O}\rangle = \frac{(-1)^k}{k!}\frac{\partial^k \mathcal{Z}}{\partial \mu^k}\bigg|_{\mu=0}$$

A more useful expansion is a Taylor expansion in $e^{-\mu}$:

$$\mathcal{Z}_\mu[\mathcal{O}] = \sum_{n=0}^{N} e^{-\mu n} P_n[\mathcal{O}] \tag{4.16}$$

in which the coefficients $P_n[\mathcal{O}]$ is the percentage of terms in $\mathcal{O}$ having size n.

4.3 Including Temperature

The main goal of the current work is to understand the role of temperature in operator growth. After all, the dynamics of the operator $\psi_1(t)$ under Heisenberg evolution has no knowledge about temperature.

One natural way to introduce temperature is to consider the operator $\rho^{1/2}$ where $\rho = Z_\beta^{-1}\exp(-\beta H)$ is thermal state at inverse temperature β. The purification of $\rho^{1/2}$ is the thermofield double (TFD) state $|TFD\rangle$ (the other factor of $\rho^{1/2}$ from the full ρ is used to make $\langle TFD|$)

$$|TFD\rangle = Z_\beta^{-1/2} e^{-\frac{\beta}{4}(H_L+H_R)}|0\rangle = Z_\beta^{-1/2} e^{-\frac{\beta}{2}H_L}|0\rangle = \left|\rho^{1/2}\right\rangle$$

where the Hamiltonians H_L, H_R are required to satisfy the condition $(H_L - H_R)|0\rangle = 0$. This state is a natural choice for studying thermodynamic properties, because for each operator $\mathcal{O}$, we can consider the corresponding operator $\mathcal{O}\rho^{1/2}$, and its average size will be directly measured by the finite temperature four-point function:

$$\delta\left[\mathcal{O}\rho^{1/2}\right] = 1 - \frac{n\left[\mathcal{O}\rho^{1/2}\right]}{N/2} = \frac{(-1)^{|\mathcal{O}|}}{N}\sum_{j=1}^{N}\mathrm{Tr}\left(\rho^{1/2}\mathcal{O}^{\dagger}\psi_j\mathcal{O}\rho^{1/2}\psi_j\right) \tag{4.17}$$

By considering the generating functions $\mathcal{Z}_\mu\left[\rho^{1/2}\right]$ and $\mathcal{Z}_\mu\left[\mathcal{O}\rho^{1/2}\right]$, we will be able to study the higher moments of size. By extracting the size distribution of $\rho^{1/2}$ from the size distribution of $\mathcal{O}\rho^{1/2}$, we find a natural notion for a "thermal" size of $\mathcal{O}$.

4.3.1 Thermal State

We begin by studying $\rho^{1/2}$. Taking $\mathcal{O} = \rho^{1/2}$ in Eq. (4.14), we find the following relation between the thermal two-point function and the size operator

$$G\left(\frac{\beta}{2}\right) = \frac{1}{N}\sum_{j=1}^{N} Z_\beta^{-1}\mathrm{Tr}\left(e^{-\beta H}\psi_j\left(\frac{\beta}{2}\right)\psi_j\right) = 1 - \frac{n\left[\rho^{1/2}\right]}{N/2} = \delta\left[\rho^{1/2}\right]$$

This relation tells us that the most de-correlated value of the Euclidean two-point function - $G(\beta/2)$ - is equal to the fractional distance the operator $\rho^{1/2}$ is from being scrambled

$$\delta_\beta \equiv \delta\left[\rho^{1/2}\right] = G\left(\frac{\beta}{2}\right) \tag{4.18}$$

which implies that the average size of $\rho^{1/2}$ is given by

$$n\left[\rho^{1/2}\right] = \frac{N}{2}\left(1 - G\left(\frac{\beta}{2}\right)\right) \tag{4.19}$$

In the high temperature limit $\beta \to 0$, one expects $G(\beta/2) \simeq G(0) = 1$, since the fermions square to one (4.5), which is consistent with the fact that $\rho^{1/2}$ approaches identity and the size shrinks to zero. On the contrary, in the low temperature limit $\beta \to \infty$, if $G(\beta/2) \to 0$, the size of $\rho^{1/2}$ approaches the scrambled (typical) value $N/2$. This result is very general since the imaginary time two-point function $G(\tau)$ decays in most physical systems. For example, in all systems with a unique ground state and an excitation gap, $G(\tau)$ decays exponentially at low temperature limit, so that $G(\beta/2) \to 0$ when $\beta \to \infty$. In a conformal field theory, $G(\tau)$ decays in power law in the zero temperature limit, which also leads to the same length $n_{\beta\to\infty} = N/2$.

To learn more than just the average size, we construct the generating function

$$\mathcal{Z}_\mu\left[\rho^{1/2}\right] = \left\langle \rho^{1/2}\right| e^{-\mu n}\left|\rho^{1/2}\right\rangle = \langle TFD|e^{-\mu n}|TFD\rangle \tag{4.20}$$

Therefore learning about the operator distribution of $\rho^{1/2}$ is equivalent to learning about the fermion number distribution in the thermofield double state.

4.3.2 Thermal Fermion

If we take $\mathcal{O} = \psi_1(t)$ in Eq. (4.17), we see that the average size of $\psi(t)\rho^{1/2}$ is entirely equivalent to an out-of-time-order correlator (a.k.a. OTOC) [21,94,96,97]:

$$-\frac{1}{N}\sum_{j=1}^{N}\mathrm{Tr}\left(\rho^{1/2}\psi_1(t)\psi_j\psi_1(t)\rho^{1/2}\psi_j\right) = 1 - \frac{2}{N}n\left[\psi_1(t)\rho^{1/2}\right] \equiv \delta\left[\psi_1(t)\rho^{1/2}\right]$$

Therefore, the statement that the OTOC de-correlates exponentially is equivalent to the statement that the average size of the operator $\psi_1(t)\rho^{1/2}$ grows exponentially. If the OTOC vanishes in long time, that implies that the size of $\psi_1(t)\rho^{1/2}$ reaches the scrambled value $n_* = N/2$.

The size distribution of $\psi_1(t)\rho^{1/2}$ can be uncovered through the generating function

$$\mathcal{Z}_\mu\left[\psi_1(t)\rho^{1/2}\right] = \langle\psi_1(t)\rho^{1/2}|e^{-\mu n}|\psi_1(t)\rho^{1/2}\rangle = \langle TFD|\psi_1^L(t)e^{-\mu n}\psi_1^L(t)|TFD\rangle \tag{4.21}$$

The operator $e^{-\mu n}$ can be viewed as an Euclidean time evolution with time μ and Hamiltonian n, so that the generating function (4.21) is related to the two-point function in a system with time-dependent Euclidean evolution:

$$\mathcal{G}_\mu(\tau_a,\tau_b) = \frac{\langle 0|\,\mathcal{T}\left[e^{-\beta(H_L+H_R)/2}e^{-\mu n(\beta/4)}\psi_1^L(\tau_a)\psi_1^L(\tau_b)\right]|0\rangle}{\langle 0|\,\mathcal{T}\left[e^{-\beta(H_L+H_R)/2}e^{-\mu n(\beta/4)}\right]|0\rangle} \tag{4.22}$$

where $\mathcal{T}$ is the Euclidean time ordering symbol and $\psi_1^L(\tau_{a,b})$ are imaginary time evolved fermion operators. Note that the denominator in (4.22) is, up to a factor of thermal partition function Z_β that cancels with the numerator, exactly the

size generating function $\mathcal{Z}_\mu\left[\rho^{1/2}\right]$ in Eq. (4.20). Therefore, the size generating function $\mathcal{Z}_\mu\left[\psi_1(t)\rho^{1/2}\right]$ naturally factorizes into the product of

$$\mathcal{Z}_\mu\left[\psi_1(t)\rho^{1/2}\right] = \mathcal{G}_\mu\left(\frac{\beta^+}{4}+it, \frac{\beta^-}{4}+it\right)\mathcal{Z}_\mu\left[\rho^{1/2}\right] \tag{4.23}$$

This equation clarifies that the two-point function $\mathcal{G}_\mu\left(\beta/4^+ + it, \beta/4^- + it\right)$ measures the size change $\psi_1(t)$ induces upon $\rho^{1/2}$ through multiplication. We can see this directly by applying the expansion in Eq. (4.16) to both sides of this equation, in order to obtain the following convolution formula for the size distribution of $\psi_1\left(t\right)\rho^{1/2}$

$$P_n\left[\psi_1(t)\rho^{1/2}\right] = \left(K^\beta\left[\psi_1\left(t\right)\right] * P\left[\rho^{1/2}\right]\right)_n = \sum_{m=0}^{n} K_m^\beta\left[\psi_1(t)\right] P_{n-m}\left[\rho^{1/2}\right] \tag{4.24}$$

with $K_m^\beta\left[\psi_1(t)\right]$ defined by the expansion of $\mathcal{G}_\mu\left(\beta/4^+ + it, \beta/4^- + it\right)$ in powers of $e^{-\mu}$:

$$\mathcal{G}_\mu\left(\frac{\beta^+}{4}+it, \frac{\beta^-}{4}+it\right) = \sum_{m=0}^{N} e^{-m\mu} K_m^\beta\left[\psi_1(t)\right] \tag{4.25}$$

In this sense, K_m^β can be viewed as the "growth distribution" caused by applying $\psi_1(t)$ to the thermal state $\rho^{1/2}$.

Note that the discussion above can be generalized to arbitrary operators. For an arbitrary operator $\mathcal{O}$, as long as we normalize it such that $\langle TFD|\mathcal{O}_L^\dagger\mathcal{O}_L|TFD\rangle \equiv \left\langle\mathcal{O}^\dagger\mathcal{O}\right\rangle_\beta = 1$, the expansion of the two-point function

$$\mathcal{G}_\mu\left[\mathcal{O}\right] \equiv \frac{\mathcal{Z}_\mu\left[\mathcal{O}\rho^{1/2}\right]}{\mathcal{Z}_\mu\left[\rho^{1/2}\right]} \tag{4.26}$$

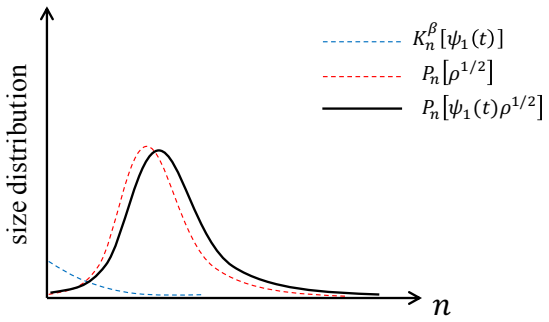


Figure 4.3: Schematic illustration of the size distribution $P_n\left[\psi_1\left(t\right)\rho^{1/2}\right]$ for the operator $\psi_1\left(t\right)\rho^{1/2}$ (black curve) which naturally decomposes into a convolution of a growth distribution $K_n^{\beta}\left[\psi_1\left(t\right)\right]$ (blue dashed curve) with the size distribution $P_n\left[\rho^{1/2}\right]$ of the operator $\rho^{1/2}$ (red dashed curve). This is due to the factorization relation of their respective generating functions (4.23).

measures the effective size distribution of $\mathcal{O}$ when applied to the thermal state.

4.3.3 Twisted Boundary Condition

We are interested in studying the generating function $\mathcal{Z}_{\mu}\left[\mathcal{O}\right]$ for $\mathcal{O}=\rho^{1/2}$ and $\mathcal{O}=\psi_1(t)\rho^{1/2}$. As we discussed earlier, inserting the operator $\exp\left(-\mu n\right)$ corresponds to changing the imaginary time evolution. The computation can be simplified by noticing that $\exp\left(-\mu n\right)$ is a Gaussian operator, such that its action by conjugation to fermion operators $\psi_i^{L,R}$ leads to a simple linear superposition:

$$e^{\mu n}\begin{pmatrix}\psi^L\\ i\psi^R\end{pmatrix}e^{-\mu n}=\begin{pmatrix}\cosh\left(\mu\right) & -\sinh\left(\mu\right)\\ -\sinh\left(\mu\right) & \cosh\left(\mu\right)\end{pmatrix}\begin{pmatrix}\psi^L\\ i\psi^R\end{pmatrix} \tag{4.27}$$

As a result, inserting the operator-weighting term $\exp\left(-\mu n\right)$ is equivalent to twisting the boundary condition of the fermion fields at $\tau=\beta/4$.

It is convenient to "de-purify" the system and return to the single copy of

fermion fields, but with a twisted boundary condition. The single field is defined by continuously stitching the left and right fields together:

$$\psi_i(\tau) = \begin{cases} \psi_i^L(\tau) & 0 \le \tau \le \beta/2 \\ i\psi_i^R(\beta - \tau) & \beta/2 \le \tau < \beta \end{cases} \tag{4.28}$$

with the requirement of course that $\psi(\tau + \beta) = -\psi(\tau)$. This stitching transforms the purified action for the two fields into the original action for this single field; however, the twist condition must accompany the fields. In conclusion, the two-sided path integral in the presence of the factor $\exp(-\mu n(\beta/4))$ equals the original path integral where the fields are twisted according to

$$\lim_{\tau \to \beta/4^+} \begin{pmatrix} \psi(\tau) \\ \psi(\beta - \tau) \end{pmatrix} = \begin{pmatrix} \cosh(\mu) & -\sinh(\mu) \\ -\sinh(\mu) & \cosh(\mu) \end{pmatrix} \lim_{\tau \to \beta/4^-} \begin{pmatrix} \psi(\tau) \\ \psi(\beta - \tau) \end{pmatrix} \tag{4.29}$$

Therefore, we conclude that calculating the two point function $\mathcal{G}_\mu$ is equivalent to calculating the original two-point function, but with the following twisted boundary conditions

$$\begin{pmatrix} \lim_{\tau_{1/2} \to \beta/4^+} \mathcal{G}_\mu(\tau_1, \tau_2) \\ \lim_{\tau_{1/2} \to 3\beta/4^-} \mathcal{G}_\mu(\tau_1, \tau_2) \end{pmatrix} = \begin{pmatrix} \cosh(\mu) & -\sinh(\mu) \\ -\sinh(\mu) & \cosh(\mu) \end{pmatrix} \begin{pmatrix} \lim_{\tau_{1/2} \to \beta/4^-} \mathcal{G}_\mu(\tau_1, \tau_2) \\ \lim_{\tau_{1/2} \to 3\beta/4^+} \mathcal{G}_\mu(\tau_1, \tau_2) \end{pmatrix} \tag{4.30}$$

We note that while these conditions break time translation invariance, they preserve a set of discrete symmetries. Specifically, if the original Hamiltonian is

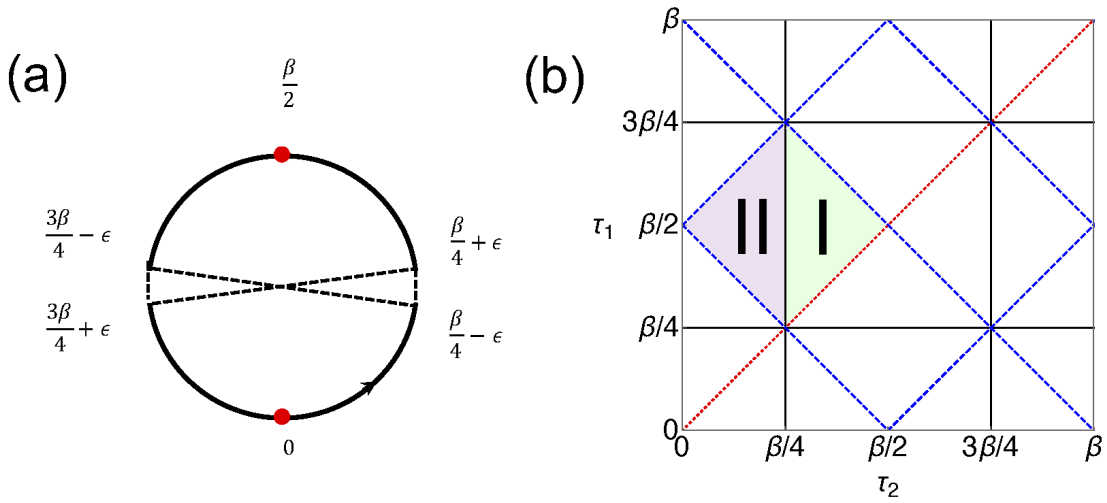


Figure 4.4: (a) The twisted boundary condition on the imaginary time circle. When τ crosses $\beta/4$ from below, $\psi(\tau)$ becomes a superposition of $\psi(\beta/4+\epsilon)$ and $\psi(3\beta/4-\epsilon)$ (see Eq. (4.29). (b) The various symmetry and boundary conditions on the twisted two point function in the (τ_1, τ_2) plane. First, $\mathcal{G}_\mu$ is odd under reflections across the red dotted line and even under reflections across the blue dashed lines. Thus, it is sufficient to solve the saddle-point equations in the fundamental domain $0 < \tau_1 - \tau_2 < \beta/2$ and $\beta/2 < \tau_1 + \tau_2 < \beta$. The black lines are the locations of the twisting boundary conditions (4.30), which reduce in the large q limit (4.38) and divide our fundamental domain into two regions. Region I is where neither of the two fermions have crossed a twist, while in Region II the fermions are on opposite sides of the twist.

time-reversal invariant, then $\mathcal{G}_\mu(\tau_1, \tau_2)$ has reflection symmetry across the lines $\tau_1 \pm \tau_2 = n\beta/2$ for all integers $n \in \mathbb{Z}$

$$\mathcal{G}_\mu(\tau_1, \tau_2) = \mathcal{G}_\mu\left(\frac{n\beta}{2} - \tau_2, \frac{n\beta}{2} - \tau_1\right) = (-1)^{n+1}\,\mathcal{G}_\mu\left(\tau_2 + \frac{n\beta}{2}, \tau_1 - \frac{n\beta}{2}\right) \quad (4.31)$$

Thus, we need only to solve for $\mathcal{G}_\mu(\tau_1, \tau_2)$ in the fundamental domain $0 < \tau_1 - \tau_2 < \beta/2$ and $\beta/2 < \tau_1 + \tau_2 < \beta$, as shown in Fig. 4.4(b) by the union of regions I and II.

4.3.4 Thermally Renormalized Unit of Size

As an interesting application of our formalism, let us note how $\psi_1(t)$ affects $\rho^{1/2}$ by taking $t = 0$ and consider the change of average size by a single fermion operator ψ_1.

$$\begin{aligned}\Delta n_\beta[\psi_1] &\equiv n\left[\psi_1\rho^{1/2}\right] - n\left[\rho^{1/2}\right] \\ &= \frac{1}{2}\sum_{i=1}^{N}\left(\langle TFD|\,\psi_1^L i\psi_i^L\psi_i^R\psi_1^L\,|TFD\rangle - \langle TFD|\,i\psi_i^L\psi_i^R\,|TFD\rangle\right) \\ &= \langle TFD|\,i\psi_1^R\psi_1^L\,|TFD\rangle = G_{11}\left(\frac{\beta}{2}\right)\end{aligned} \tag{4.32}$$

At infinite temperature $\beta \to 0$, $G(\beta/2) = 1$, which restores the trivial result that ψ_1 increases the size of the density operator (which is proportional to identity operator, with size 0) by 1. At finite temperature, interestingly, the size change induced by a single fermion operator is smaller than 1, and is given by the same imaginary time two-point function as the one that determines the fractional scrambling distance $\delta_\beta = 1 - \frac{n\left[\rho^{1/2}\right]}{N/2}$ in Eq. (4.18). In general, the size increase induced by ψ_i is $\Delta n_\beta[\psi_i] = G_{ii}(\beta/2)$, which may depend on i. The average size increase is exactly δ_β. [2]

$$\frac{1}{N}\sum_i \Delta n[\psi_i] = \frac{1}{N}\sum_i G_{ii}\left(\frac{\beta}{2}\right) = G\left(\frac{\beta}{2}\right) = \delta_\beta \tag{4.33}$$

Physically, the average size change due to applying a single fermion is generically $\delta_\beta < 1$ at finite β, because in the presence of a nontrivial $\rho^{1/2}$ there is a chance that multiplying by ψ_1 decreases the size, as is illustrated in Fig. 4.5, although

[2] In term of the probabilities $K_m^\beta[\psi_1]$ in Eq. (4.25), we have $\Delta n[\psi_i] = \sum_{m=0}^{N} m K_m^\beta[\psi_1]$.

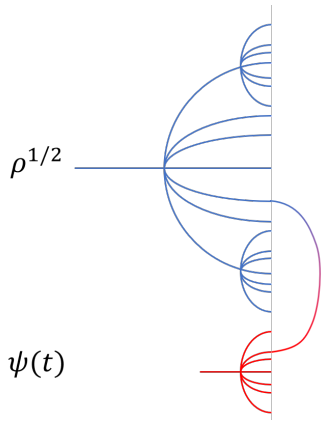


Figure 4.5: At finite temperature, when an operator such as $\psi_1(t)$ is multiplied to $\rho^{1/2}$, there is a chance that some fermion flavors collide and the size increase is smaller than the size of $\psi_1(t)$ itself.

the chance of increasing the size is always bigger. The closer the length of $\rho^{1/2}$ is to the scrambling value $n_* = N/2$, the smaller is the size increase $\Delta n_\beta \left[\psi_1 \rho^{1/2}\right]$. For a fully scrambled operator with $n = N/2,\ \delta = 0$, multiplying a fermion ψ_1 has equal chance of increasing or decreasing the size, so that the average size stays the same.

It should be emphasized that the discussion above is not restricted to the thermal density operator. For any operator $\mathcal{O}$, we can define the size change

$$\Delta n_{\mathcal{O}}\left[\psi_i\right] \equiv n\left[\psi_i \mathcal{O}\right] - n\left[\mathcal{O}\right]$$

and obtain the following identity:

$$\frac{1}{N}\sum_i \Delta n_{\mathcal{O}}\left[\psi_i\right] = \delta\left[\mathcal{O}\right] \equiv 1 - \frac{2n\left[\mathcal{O}\right]}{N}$$

The only thing special for the thermal density operator is the relation of Δn to

imaginary time two-point function in a single-copy system.

Furthermore, instead of ψ_i we can consider a string

$$\Gamma_I \equiv \Gamma_{i_1 i_2 ... i_k} = i^{\frac{k(k-1)}{2}} \psi_{i_1} ... \psi_{i_k} \qquad 1 \leq i_1 < i_2 < ... < i_k \leq N$$

introduced in Eq. (4.6), and consider how the size of $\Gamma_I \mathcal{O}$ is different from $\mathcal{O}$. We have

$$\Delta n_{\mathcal{O}} [\Gamma_I] \equiv n [\Gamma_I \mathcal{O}] - n [\mathcal{O}] = \sum_{s=1}^{k} \langle \mathcal{O}| i \psi_{i_s}^R \psi_{i_s}^L |\mathcal{O}\rangle \tag{4.34}$$

If we average over all Majorana strings Γ_I with the same size k, we obtain

$$\frac{1}{C_N^k} \sum_I \Delta n_{\mathcal{O}} [\Gamma_I] = k\delta [\mathcal{O}] \tag{4.35}$$

In the last equation, $C_N^k = \frac{N!}{k!(N-k)!}$ is the number of strings with length k. This equation shows that the average size change induced by multiplying a string with length k is k times $\delta [\mathcal{O}]$, further confirms that each fermion in the string contributes additively.

This observation suggests that at finite temperature (or more generally, for any density operator ρ), the fractional scrambling distance δ, rather than 1, should be considered as the fundamental unit of size, which is carried by each fermion operator. Indeed, as we will discuss in next section, our calculation in the SYK model in the large q limit suggests universal behavior occurs when size is measured in this renormalized unit.

4.4 SYK Model

In this section, we will study the operator size growth in the SYK model [4,98]. This model features q-local interactions with independently random couplings, where each of the couplings is normal distributed

$$H = i^{q/2} \sum_{1 \leq i_1 \ldots \leq i_q \leq N} J_{i_1 \ldots i_q} \psi_{i_1} \ldots \psi_{i_q} \qquad \left\langle J^2_{i_1 \ldots i_q} \right\rangle = \frac{J^2}{\binom{N-1}{q-1}} = \frac{\mathcal{J}^2}{2q\binom{N-1}{q-1}}$$

with $\{\psi_i, \psi_j\} = 2\delta_{ij}$. At large N, the two-point function satisfies the saddle-point equations

$$[G]^{-1} = [G_0]^{-1} - [\Sigma] \qquad \Sigma(\tau_1, \tau_2) = \frac{\mathcal{J}^2}{2q} \left(G(\tau_1, \tau_2)\right)^{q-1} \tag{4.36}$$

where bracketed terms are Matsubara frequency matrices. One should note that since the fermions square to one, $[G_0]^{-1} = -i\omega/2$ rather than $-i\omega$.

4.4.1 Large q Approximation

In the language of Feynman diagrams, the Schwinger-Dyson equation (4.36) corresponds to only keeping the leading "melon diagrams" as is shown in Fig. 4.6. All other diagrams are sub-leading in large N. In the large q limit, there are two types of diagrams. Those with melons inserted into melons (such as Fig. 4.6(a)) receive a combinatorial q enhancement, as there are many rungs upon which one may insert (hence the need for a q^{-1} factor in the self-energy to keep everything finite). In contrast, diagrams where melons are simply threaded together (such as Fig. 4.6(b)) do not receive this enhancement [30]. Thus, at large q only the former

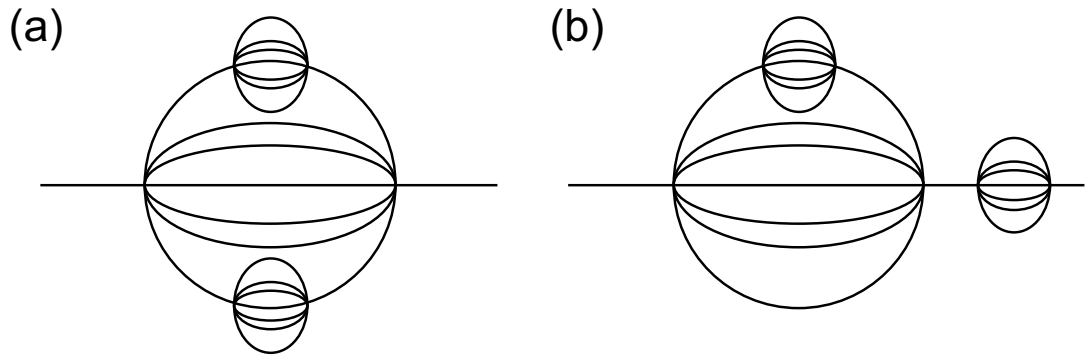


Figure 4.6: Two $\mathcal{O}\left(J^6\right)$ examples of the planar graphs of that survive the large-N limit. Only graphs of form (a), where melons are inserted into melons, survive the large q limit. Notice that there are $\mathcal{O}\left(q^2\right)$ graphs of form (a), while there are only $\mathcal{O}\left(q\right)$ graphs of form (b). This is because there are $\mathcal{O}\left(q\right)$ locations at any depth of a given graph to insert another melon; however, there are typically only $\mathcal{O}\left(1\right)$ locations to thread another melon. Accordingly, we take our coupling J to equal $\mathcal{J}/\sqrt{2q}$. Consequently, the $\mathcal{O}\left(q\right)$ combinatorial enhancement gained for each new melon insertion is canceled by the $J^2 = \mathcal{J}^2/\left(2q\right)$ factor accompanying said melon. This q-scaling of the coupling isolates the infinite subset of the planar graphs where the graphs are two copies of a tree that are then glued together (a.k.a. "doubletree" graphs) such as (a). All non-doubletree graphs such as (b) are suppressed in q since they receive factors of $\mathcal{J}^2/\left(2q\right)$ for each melon, but do not receive the necessary number of q combinatorial enhancements.

dominate, which corresponds to the following truncation of the Schwinger-Dyson expansion:

$$[G] = [G_0] + [G_0]\,[\Sigma]\,[G_0] \qquad \Sigma\left(\tau_1, \tau_2\right) = \frac{\mathcal{J}^2}{2q}\left(G\left(\tau_1, \tau_2\right)\right)^{q-1}$$

Combing the equations together and Fourier transforming, one obtains

$$\partial_{\tau_1}\partial_{\tau_2}\left(G - G_0\right) = -\frac{2\mathcal{J}^2}{q}G^{q-1}$$

The role played by G_0 in this equation is to require that $G \to G_0$ as τ_{12} goes to

integer multiples of β. Therefore, if we take $G = G_0 e^{\sigma/q}$ with $\sigma \to 0$ at the τ_{12} boundaries, we obtain Liouville's equation [2, 112]

$$\partial_{\tau_1}\partial_{\tau_2}\sigma = -2\mathcal{J}^2 e^{\sigma} + \mathcal{O}(1/q) \tag{4.37}$$

where the field σ is expected to be periodic in both of its arguments, as well as have kinks when τ_{12} approaches integer multiples of β.

Now in order to find $\mathcal{G}_\mu$, we will need to solve the above equations with the twisted boundary conditions (4.30). Furthermore, our twisted two-point function $\mathcal{G}_\mu$ also satisfies the reflection conditions (4.31). Thus, we need only solve for $\mathcal{G}_\mu$ in the fundamental domain $0 < \tau_1 - \tau_2 < \beta/2$ and $\beta/2 < \tau_1 + \tau_2 < \beta$, as shown in Fig. 4.4(b) by the union of regions I and II.

4.4.2 The Large-q solution

In the large q limit, each commutator with the Hamiltonian increases the size of operator by $\sim q$, so that it is natural to meaure the operator size in unit of q. In the generating function, this corresponds to defining $\mu = \frac{\widehat{\mu}}{q}$, with $\widehat{\mu}$ kept finite in the large q limit. The derivative of the generating function over $\widehat{\mu}$ measures the size n in unit of q. If we consider the large q limit with $\widehat{\mu}$ being kept finite, and use the large-q ansatz for twisted two-point function

$$\mathcal{G}_\mu(\tau_1, \tau_2) = G_0(\tau_1, \tau_2)\, e^{\sigma_\mu(\tau_1,\tau_2)/q},$$

the boundary condition (4.30) reduces to

$$\begin{aligned}\begin{pmatrix} \lim\limits_{\tau_{1/2}\to\beta/4^+}\mathcal{G}_\mu(\tau_1,\tau_2) \\ \lim\limits_{\tau_{1/2}\to 3\beta/4^-}\mathcal{G}_\mu(\tau_1,\tau_2) \end{pmatrix} &\simeq \begin{pmatrix} 1 & -\frac{\widehat{\mu}}{q} \\ -\frac{\widehat{\mu}}{q} & 1 \end{pmatrix} \begin{pmatrix} \lim\limits_{\tau_{1/2}\to\beta/4^-}\mathcal{G}_\mu(\tau_1,\tau_2) \\ \lim\limits_{\tau_{1/2}\to 3\beta/4^+}\mathcal{G}_\mu(\tau_1,\tau_2) \end{pmatrix} \\ &\simeq e^{-\widehat{\mu}/q} \begin{pmatrix} \lim\limits_{\tau_{1/2}\to\beta/4^-}\mathcal{G}_\mu(\tau_1,\tau_2) \\ \lim\limits_{\tau_{1/2}\to 3\beta/4^+}\mathcal{G}_\mu(\tau_1,\tau_2) \end{pmatrix}\end{aligned} \tag{4.38}$$

To the leading order of $\frac{1}{q}$, the two equations for $\beta/4$ and $3\beta/4$ decouple.

The twisted two-point function in large-q limit can thus be obtained by solving Liouville's equation (4.37) with the μ-dependent boundary conditions (4.38). Here we will skip the tedious details and directly present the solution. When the times are such that the two fermions are on the same side of the twisted boundary, which corresponds to $\tau_2 > \beta/4$ (region I in Fig. 4.4(b), we find a seemingly time-translation invariant solution solution

$$\mathcal{G}_\mu(\tau_1,\tau_2) = \left(\frac{\sin\gamma_\mu}{\sin\left(\alpha_\mu(\tau_1-\tau_2)+\gamma_\mu\right)}\right)^{2/q} \equiv G_\mu(\tau_1-\tau_2) \tag{4.39}$$

However, when the times are such that the two fermions are on opposite sides of the twisted boundary at $\beta/4$ and $3\beta/4$, which for our domain amounts to the condition $\tau_2 < \beta/4$ (region II in Fig. 4.4(b), the time translation symmetry is

explicitly broken

$$\mathcal{G}_\mu(\tau_1,\tau_2)=\frac{e^{-\widehat{\mu}/q}G_\mu(\tau_1-\tau_2)}{\left(1-\frac{\left(1-e^{-\widehat{\mu}}\right)}{\sin^2\gamma_\mu}\left(G_\mu(\tau_1-\tau_2)\right)^{q/2}\sin\left(\alpha_\mu\left(\tau_1-\frac{\beta}{4}\right)\right)\sin\left(\alpha_\mu\left(\tau_2-\frac{\beta}{4}\right)\right)\right)^{2/q}} \tag{4.40}$$

Here the parameters α_μ and γ_μ are functions of $\beta\mathcal{J}$ and μ, which are determined by the boundary condition $\mathcal{G}(\tau,\tau)=1$ as well as the reflection conditions (4.31)

$$\alpha_\mu\beta=\beta\mathcal{J}\sin\gamma_\mu,\quad \sin\left(\frac{\alpha_\mu\beta}{2}+2\gamma_\mu\right)=e^{-\widehat{\mu}}\sin\left(\frac{\alpha_\mu\beta}{2}\right) \tag{4.41}$$

In the limit $\mu\to 0$, we recover the untwisted two-point function $\mathcal{G}_{\mu=0}(\tau_1,\tau_2)=G_{\mu=0}(\tau_1-\tau_2)=G(\tau_1-\tau_2)$ in the whole domain, and the equation for the parameters reduce to the ordinary case [2]:

$$\alpha_{\mu=0}\equiv\alpha=\mathcal{J}\cos\left(\frac{\alpha\beta}{2}\right),\quad \gamma_{\mu=0}\equiv\gamma=\frac{\pi-\alpha\beta}{2} \tag{4.42}$$

The asymptotic behavior at small values of $\beta\mathcal{J}$ and large values of $\beta\mathcal{J}$ respectively are given by

$$\alpha=\mathcal{J}\left(1-\frac{\beta^2\mathcal{J}^2}{8}+\mathcal{O}\left(\beta^4\mathcal{J}^4\right)\right)\qquad \alpha=\frac{\pi}{\beta}\left(1-\frac{\pi}{\beta\mathcal{J}}+\mathcal{O}\left(\frac{1}{\beta^2\mathcal{J}^2}\right)\right)$$

4.4.3 Size renormalization

Before carrying further analysis to the SYK operator growth in next section, we need to discuss an important modification to the two-point function solution

due to higher order q effects. If we take $\tau_1 \to \beta/4+\epsilon,\ \tau_2 \to \beta/4-\epsilon$ in Eq. (4.40), we obtain

$$\mathcal{G}_\mu\left(\frac{\beta}{4}+\epsilon, \frac{\beta}{4}-\epsilon\right) = e^{-\widehat{\mu}/q} \tag{4.43}$$

This is the kernel that determines the size change induced by multiplying ψ_1 to $\rho^{1/2}$, which has been discussed in section (4.3.4). Taking the μ-derivative of $\mathcal{G}_\mu$, we find

$$\Delta n_\beta\left[\psi_1\right] \equiv n\left[\psi_1\rho^{1/2}\right] - n\left[\rho^{1/2}\right] = -\,\partial_\mu \log \mathcal{G}_\mu\left(\frac{\beta}{4}+\epsilon, \frac{\beta}{4}-\epsilon\right)\Big|_{\mu=0} = 1$$

However, we also know that the size change is directly determined by the two-point function due to Eq. (4.32):

$$\Delta n_\beta\left[\psi_1\right] = \delta_\beta = G\left(\frac{\beta}{2}\right) = \left(\frac{\alpha}{\mathcal{J}}\right)^{2/q}$$

where $\alpha \equiv \alpha_{\mu=0}\left(\beta\mathcal{J}\right)$ is the smallest positive root of Eq. (4.42).

This discrepancy between the two calculations is because $\delta_\beta \to 1$ in the large q limit, and the $\mathcal{O}\left(q^{-1}\right)$ difference is neglected in the approximation we made to the boundary condition. The easiest way to resolve this issue and makes a consistent large-q limit is by redefining $\widehat{\mu} = q\mu$ to

$$\widehat{\mu} = q\mu\delta_\beta \tag{4.44}$$

in Eq. (4.38), which leads to the same substitution in Eqs. (4.39), (4.40), and (4.41). In the following, we will always use this definition of $\widehat{\mu}$.

Physically, this substitution is a consequence of the size renormalization dis-

cussed in section (4.3.4). Each Majorana fermion increase the operator size by δ_β rather than 1. Each action of the Hamiltonian increases the operator size by $\sim q\delta_\beta$. Although in large q limit $1-\delta_\beta$ is order q^{-1}, it is important to keep track of this distance, since the same δ_β also measures the fractional scrambling distance of $\rho^{1/2}$, as we discussed in section (4.3.1). The size of $\rho^{1/2}$ is

$$n\left[\rho^{1/2}\right] = \frac{N}{2}\left(1-\delta_\beta\right) = \frac{N}{2}\left(1-\left(\frac{\alpha}{\mathcal{J}}\right)^{2/q}\right)$$

which decreases with increasing q, but is always large since we should always take the large N limit before taking large q.

4.5 SYK Operator Growth

We are now equipped with everything we need to understand $P_n\left[\psi_1\left(t\right)\rho^{1/2}\right]$, the size distribution of $\psi_1\left(t\right)\rho^{1/2}$. According to Eq. (4.23), the generating function $\mathcal{Z}_\mu\left[\psi_1(t)\rho^{1/2}\right]$ for this distribution splits into a product of the thermal state's generating function $\mathcal{Z}_\mu\left[\rho^{1/2}\right]$ and $\mathcal{G}_\mu\left(\beta/4^+ + it, \beta/4^- + it\right)$. The latter is simply the twisted two-point function we discussed in the previous subsection with an analytic continuation.

4.5.1 Thermal State

The generating function $\mathcal{Z}_\mu\left[\rho^{1/2}\right]$ is the partition function of the system with the insertion $\exp\left(-\mu n\left(\beta/4\right)\right)$ divided by that of the original system (see Eq. (4.20)). This quantity can be determined by the twisted two-point function, since

one has

$$-\frac{2}{N}\partial_\mu \ln \mathcal{Z}_\mu = \frac{2}{N}\frac{\langle \rho^{1/2}|ne^{-\mu n}|\rho^{1/2}\rangle}{\langle \rho^{1/2}|e^{-\mu n}|\rho^{1/2}\rangle} = 1 - G_\mu\left(\frac{\beta}{2}\right) = 1 - \frac{\sin^{2/q}\gamma_\mu}{\sin^{2/q}\left(\alpha_\mu \frac{\beta}{2} + \gamma_\mu\right)}$$

In theory, we can integrate this equation to obtain $\mathcal{Z}_\mu\left[\rho^{1/2}\right]$. However, many important properties of the distribution can be inferred from just the first and second moment.

The first moment is simply the average size

$$n\left[\rho^{1/2}\right] = \frac{N}{2}\left(1-\delta_\beta\right) = \frac{N}{2}\left(1 - G\left(\frac{\beta}{2}\right)\right) = \frac{N}{2}\left(1 - \left(\frac{\alpha}{\mathcal{J}}\right)^{2/q}\right) \tag{4.45}$$

The behavior of $n\left[\rho^{1/2}\right]$ for two different q values are plotted in Fig. 4.7 as a function of $\beta\mathcal{J}$. Interestingly, we see that for larger q it takes larger values of $\beta\mathcal{J}$ to achieve the same average size. Thus, the exponentiation of increasingly heavy SYK Hamiltonians results in relatively *lighter* thermal states. This is unintuitive, so one might argue that it is simply due to the q-scaling nature of $\mathcal{J}$, but that is only a constant shift in the log scale plot in Fig. 4.7, which is nowhere large enough to account for the above discrepancy. The true origin of this effect is the power of $2/q$ in the two-point function. We conclude that this heavy-light relationship is thus a non-trivial consequence of the large-N and large-q limit.

For $\beta\mathcal{J} \gg e^q$, one expects that the higher order corrections to Liouville's equation (4.37) cannot be neglected, and so one should turn to the conformal approximation [2]. This leads to the following prediction for the average size of

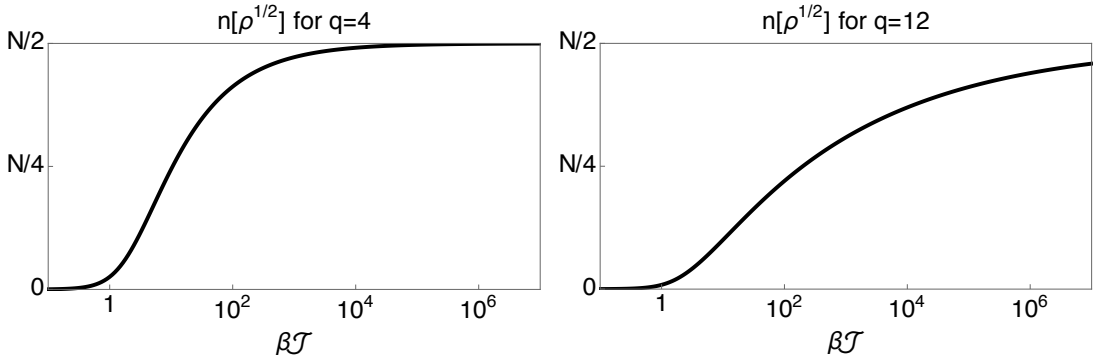


Figure 4.7: Plots of the average size of $\rho^{1/2} \propto \exp(-\beta H/2)$ for different q, given by Eq. (4.45).

$\rho^{1/2}$ when $N \gg \beta\mathcal{J} \gg e^q$

$$n\left[\rho^{1/2}\right] \approx \frac{N}{2}\left(1 - c\left(q\right)\left(\frac{\pi}{\beta\mathcal{J}}\right)^{2/q}\right) \qquad c\left(q\right) = \left(\frac{\left(q-2\right)\tan\frac{\pi}{q}}{\pi}\right)^{1/q}$$

The difference between the large q and low temperature conformal two-point functions is captured by the pre-factor $c\left(q\right)$. It monotonically increases from $\left(2/\pi\right)^{1/4} \approx 0.9$ when $q = 4$, asymptoting to 1 when q is large as $1 - 2q^{-2}$. We expect that $q = 4$ and large $\beta\mathcal{J}$ will be where our large q approximation will have the largest error. That this error is at worst 10% renews our confidence that the large q approximation captures the qualitative features of the large-N SYK model.

The second derivative of $\ln \mathcal{Z}_\mu$ determines the width of the distribution:

$$\sigma_n^2\left[\rho^{1/2}\right] = \lim_{\mu\to 0}\partial_\mu^2 \ln \mathcal{Z}_\mu\left[\rho^{1/2}\right] = \frac{N}{2}\,\partial_\mu G_\mu\left(\frac{\beta}{2}\right)\Big|_{\mu\to 0} \propto N$$

Therefore the width of the distribution $\sigma_n \propto \sqrt{N}$, such that the relative deviation

from the average value $\sigma_n\left[\rho^{1/2}\right]/n\left[\rho^{1/2}\right] \propto N^{-1/2}$ is sharply peaked in the large N limit. This is a consequence of large N factorization.

4.5.2 Thermal Fermion

As explicitly discussed in section (4.3.2), the generating function for the growth distribution K^β (4.25) is determined by the twisted two-point function (4.40)

$$\mathcal{G}_\mu\left(\frac{\beta^+}{4}+it,\frac{\beta^-}{4}+it\right)=e^{-\mu\delta_\beta}\left(1+\left(1-e^{-q\mu\delta_\beta}\right)\left(\frac{\mathcal{J}}{\alpha_\mu}\sinh\alpha_\mu t\right)^2\right)^{-2/q}$$

where α_μ and γ_μ depend on μ and $\beta\mathcal{J}$ through the constraints (4.41).

4.5.2.1 Average Size

This implies that the average size of the operator $\psi(t)\rho^{1/2}$ is given by

$$\begin{aligned} n\left[\psi_1(t)\rho^{1/2}\right] &= n\left[\rho^{1/2}\right]-\partial_\mu \ln\mathcal{G}_\mu\left(\frac{\beta^+}{4}+it,\frac{\beta^-}{4}+it\right)\Big|_{\mu=0} \\ \Rightarrow n\left[\psi_1(t)\rho^{1/2}\right] &= \frac{N}{2}(1-\delta_\beta)+\delta_\beta\left(1+2\left(\frac{\mathcal{J}}{\alpha}\sinh\alpha t\right)^2\right) \end{aligned} \tag{4.46}$$

where $\delta_\beta = (\alpha/\mathcal{J})^{2/q}$, and $\alpha \equiv \alpha_{\mu=0}(\beta\mathcal{J})$ is the smallest positive root of Eq. (4.42). We see that the difference in averages sizes of $\psi(t)\rho^{1/2}$ and $\rho^{1/2}$ is a simple when expressed in the renormalized size unit δ_β, which inspires us to define a notion of the "average growth" of $\psi_1(t)$ as

$$\Delta\tilde{n}_\beta\left[\psi_1(t)\right] \equiv \frac{n\left[\psi_1(t)\rho^{1/2}\right]-n\left[\rho^{1/2}\right]}{\delta_\beta}=1+2\left(\frac{\mathcal{J}}{\alpha}\sinh\alpha t\right)^2 \tag{4.47}$$

Now, scrambling occurs when the average size of $\psi_1(t)\,\rho^{1/2}$ given by Eq. (4.46) reaches $n_* = N/2$. This produces a slightly complicated expression for the scrambling time t_*; however, it simplifies dramatically when phrased in terms of the average growth of $\psi_1(t)$. Manipulating the scrambling time equation $n\left[\psi_1(t_*)\,\rho^{1/2}\right] = N/2$, we find that one may equivalently state that scrambling occurs when the average growth of $\psi_1(t)$ reaches $n_* = N/2$

$$\Delta\tilde{n}_\beta\left[\psi_1(t_*)\right] = 1 + 2\left(\frac{\mathcal{J}}{\alpha}\sinh\alpha t_*\right)^2 = \frac{N}{2} \tag{4.48}$$

This growth is consistent with the known result of large-q Lyapunov exponent [2].

$$\lambda_L = 2\alpha$$

4.5.2.2 Full Growth Structure

In the Lyapunov regime, we may expand the generating function of the growth distribution as

$$\mathcal{G}_\mu\left(\frac{\beta^+}{4} + it, \frac{\beta^-}{4} + it\right) = e^{-\mu\delta_\beta}\sum_{n=0}^{\infty}\binom{-2/q}{n}\left(1 - e^{-q\mu\delta_\beta}\right)^n\left(\frac{\mathcal{J}}{\alpha}\sinh\alpha t\right)^{2n}$$

where $\delta_\beta = (\alpha/\mathcal{J})^{2/q}$, and $\alpha \equiv \alpha_{\mu=0}(\beta\mathcal{J})$ is the smallest positive root of Eq. (4.42). Grouping terms by powers of $\exp(-\mu)$ and using the definition (4.25), we conclude that the growth distribution is given by

$$K^\beta_{\delta_\beta(1+qn)}\left[\psi_1(t)\right] = (-1)^n\binom{-2/q}{n}\frac{\left(\frac{\mathcal{J}}{\alpha}\sinh(\alpha t)\right)^{2n}}{\left(1 + \left(\frac{\mathcal{J}}{\alpha}\sinh(\alpha t)\right)^2\right)^{n+\frac{2}{q}}} \tag{4.49}$$

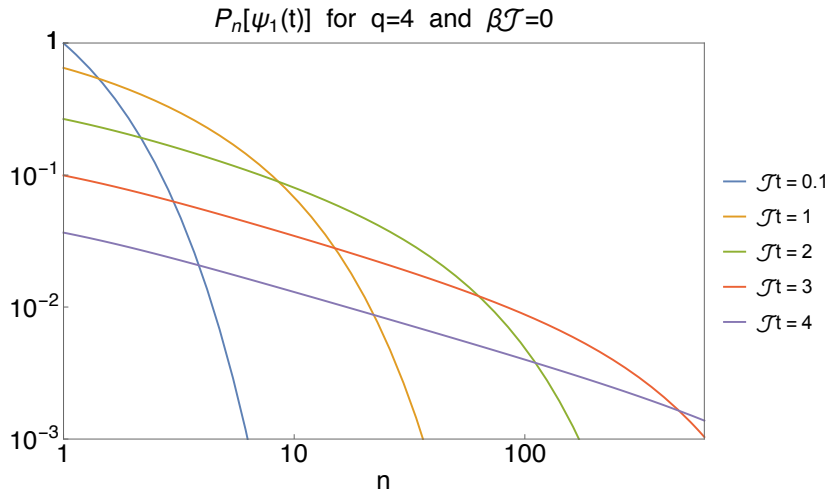


Figure 4.8: After either dynamical renormalization (4.50) or time re–parametrization (4.51), the growth distribution $K^{\beta}\left[\psi_1\left(t\right)\right]$ takes the same form as the Heisenberg evolution of the operator $\psi_1\left(t\right)$ (i.e. the infinite temperature size distribution $P\left[\psi_1\left(t\right)\right]$). This distribution is given by Eq. (4.52), and we plot it on a log-log scale. Note that it reaches out towards larger operators exponentially quickly.

where we note that $(-1)^n \binom{-2/q}{n}$ is always positive for integer n. Thus, $K^{\beta}\left[\psi_1\left(t\right)\right] \geq 0$ and so we have no negative probabilities in the size distribution of $\psi_1\left(t\right)\rho^{1/2}$, since it is given by $P\left[\psi_1\left(t\right)\rho^{1/2}\right] = K^{\beta}\left[\psi_1\left(t\right)\right] * P\left[\rho^{1/2}\right]$ as shown in section (4.3.2).

Interestingly, we see that the growth distribution K_{β} (4.49) has a functional form independent of temperature, which we plot in Fig. (4.8). We can use either of two methods to expose this phenomenon. One option is to replace the coupling $\mathcal{J}$ with the dynamically renormalized coupling $\tilde{\mathcal{J}}\left(t\right)$ (plotted in Fig. (4.9)):

$$\tilde{\mathcal{J}}(t) = \frac{\operatorname{arcsinh}\left(\frac{\mathcal{J}}{\alpha}\sinh\left(\alpha t\right)\right)}{t} = \alpha + \frac{\log\left(\mathcal{J}/\alpha\right)}{t} + \frac{\mathcal{O}\left(e^{-2\alpha t}\right)}{t} \tag{4.50}$$

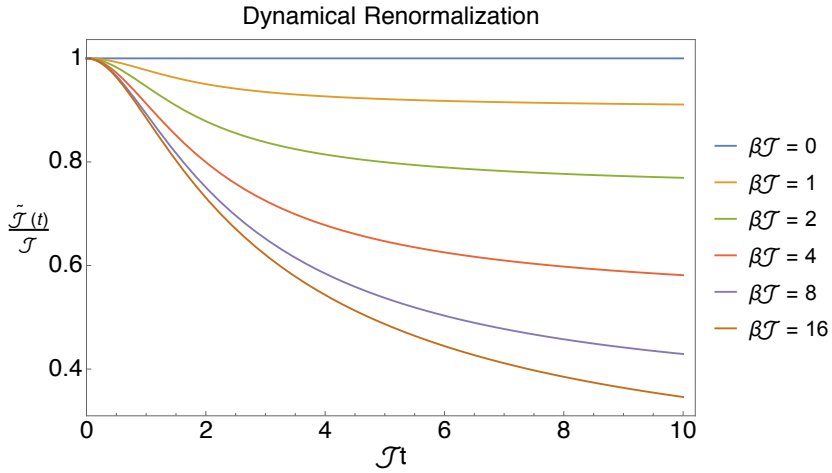


Figure 4.9: For different values of $\beta\mathcal{J}$, the effective coupling $\tilde{\mathcal{J}}(t)$ given by Eq. (4.50) slows down from $\mathcal{J}$ to α on a timescale of order α^{-1}. As always, α is the smallest root of Eq. (4.42).

The other option is to re-parametrize time

$$\tilde{t} = \frac{1}{\mathcal{J}}\text{arcsinh}\left(\frac{\mathcal{J}}{\alpha}\sinh(\alpha t)\right) = \frac{\alpha}{\mathcal{J}}t + \frac{\log(\mathcal{J}/\alpha)}{\mathcal{J}} + \frac{\mathcal{O}(e^{-2\alpha t})}{\mathcal{J}} \tag{4.51}$$

Both methods transform the finite temperature growth distribution into that of the Heisenberg evolution of the operator $\psi_1(t)$ (i.e. the infinite temperature size distribution $P_{1+qn}[\psi_1(t)]$) [7]. For example, using $\tilde{t}$ gives

$$K^{\beta}_{\delta_\beta(1+qn)}(\tilde{t}) = (-1)^n \binom{-2/q}{n} \frac{\tanh\left(\mathcal{J}\tilde{t}\right)^{2n}}{\cosh\left(\mathcal{J}\tilde{t}\right)^{\frac{4}{q}}} = K^{\beta=0}_{1+qn}\left[\psi_1\left(\tilde{t}\right)\right] = P_{1+qn}\left[\psi_1\left(\tilde{t}\right)\right] \tag{4.52}$$

This temperature-independence is fascinating since the Heisenberg evolution of $\psi_1(t)$ was obtained in [7] via fully-dressed Feynman graph calculations. In other words, the distribution $P[\psi_1(t)]$ represents the simple tree graphs such as Fig. 4.6(a) constructed using the original SYK Hamiltonian. However, we just showed

how the growth distribution $K^{\beta}\left[\psi_1\left(t\right)\right]$ can be easily transformed to $P\left[\psi_1\left(t\right)\right]$. Therefore, since $P\left[\psi_1\left(t\right)\rho^{1/2}\right] = K^{\beta}\left[\psi_1\left(t\right)\right] * P\left[\rho^{1/2}\right]$ and $P\left[\rho^{1/2}\right]$ is well-peaked, we are led to the remarkable conclusion the growth dynamics of large-N, large-q SYK model is totally universal. In fact, if one waits an initial period α^{-1} to enter the Lyapunov regime, then one need simply use the effective size δ_{β} and coupling $\tilde{\mathcal{J}} = \alpha$ for the full growth structure of $\psi_1\left(t\right)\rho^{1/2}$ to match that of $\psi_1\left(t\right)$.

4.5.3 Finite Temperature Epidemic Model

In this subsection we will discuss the physical interpretation of the SYK operator growth by relating it to an epidemic model. Intuition for operator scrambling behavior has been developed by various authors [**?**, 7, 63, 83], resulting in an infection picture for operator growth. An operator such as $\psi_1(t)$ can be expanded in the strings of Majorana fermion Γ_I. We consider the fermions already included in the string as "infected". Heisenberg evolution of Γ_I generates a term $[\Gamma_I, H]$ which could contain a few more fermions. For example for SYK model with q-body interactions, in the large N limit most of the terms have one fermion replaced by $q-1$ other fermions. In order for these $q-1$ fermions to be "infected", they must not be already in Γ_I. Therefore the infection rate depends on the infectable population.

In the simplest infection model for a population of n_* individuals, the rate of infection is proportional to the number of unexposed people times the number of contagious people

$$\frac{dn\left(t\right)}{dt} = r\left(1 - \frac{n\left(t\right)}{n_*}\right) n\left(t\right) \tag{4.53}$$

More generally, in various quantum circuit and Hamiltonian systems, both terms on the right-side of the equation may be raised to various powers or there may even be a sum of such terms, due to the potential multi-body nature of the interaction. For example, in SYK, upon a single commutation with the Hamiltonian, a size 1 operator becomes a size $q-1$ operator, so we might expect various powers of q to appear in the above expression. Regardless, in either case sigmoidal behavior will be produced, which is consistent with general expectations of four-point functions.

Let us see just how well such a picture can apply to the SYK model. Taking the derivative of Eq. (4.46) and using Eq. (4.48), we find that during the Lyapunov regime ($\log N \gg \alpha t \gg 1$)

$$\frac{d}{dt}\left(n\left[\psi_1(t)\,\rho^{1/2}\right]\right) \approx (2\mathcal{J})\left(1-\frac{n\left[\rho^{1/2}\right]}{n_*}\right)^{q/2}\left(n\left[\psi_1(t)\,\rho^{1/2}\right]-n\left[\rho^{1/2}\right]\right)$$

Comparing with the infection equation (4.53), we have the fundamental rate $r = 2\mathcal{J}$ as well one of the terms being raised to $q/2$ due to the q-local nature of the interaction. However, rather than $\left(1-n\left[\psi(t)\,\rho^{1/2}\right]/n_*\right)^{q/2}$, which one may have expected by direct analogy with the infection equation, we have the static term $\left(1-n\left[\rho^{1/2}\right]/n_*\right)^{q/2} = \delta_\beta^{q/2}$. During the Lyapunov regime, these two are the same to leading order in N. Lastly, it appears through the final term that of the large population $n\left[\psi_1(t)\,\rho^{1/2}\right]$, only the small population $n\left[\psi_1(t)\,\rho^{1/2}\right] - n\left[\rho^{1/2}\right]$ possesses the ability to infect others. Notice that there remains the large population $n\left[\rho^{1/2}\right]$ who count as having been exposed, but do not infect others. It is thus natural to view this group as a vaccinated population.

In other words, after waiting for the dynamical renormalization to settle down,

the physics of the four-point function is well-described by an infection model, with the caveat that only a small population $n\left[\psi\left(t\right)\rho^{1/2}\right]-n\left[\rho^{1/2}\right]$ possess the ability to infect. In this sense, the operator $\rho^{1/2}$ vaccinates a finite fraction of the N flavors. Now regardless of whether any particular individual possess the ability to infect, it remains that a large portion of the population has been exposed, and thus the probability for any contagious individual to encounter an unexposed individual is decreased. Consequently, the overall rate of infection slows down to

$$\lambda_L = 2\mathcal{J}\left(1-\frac{n\left[\rho^{1/2}\right]}{N_*}\right)^{q/2} = 2\mathcal{J}\delta_\beta^{q/2} = 2\mathcal{J}\left(G\left(\frac{\beta}{2}\right)\right)^{q/2} = 2\alpha$$

as illustrated in Fig. (4.10).

4.6 Discussion

The methodology developed in sections (4.2) and (4.3) is very powerful, as it applies to all fermionic systems. Specifically, determining the system's full growth distribution amounts to calculating the twisted (4.30) two-point function $\mathcal{G}_\mu$ followed by inverse transforming in μ (4.25). The large-N saddle point technique and the large-q simplification enabled us to obtain a closed solution in SYK. Even if analytics are too difficult, this analysis can be effectively implemented numerically for many classes of models.

These techniques also allowed us to compute a four-point function, since Eq. (4.17) shows that the average size of the operator $\psi\left(t\right)\rho^{1/2}$ (4.46) gives the value of a certain four-point function. We can generalize and calculate arbitrary four-point functions by moving the twist (4.30) to other locations. This has the non-trivial

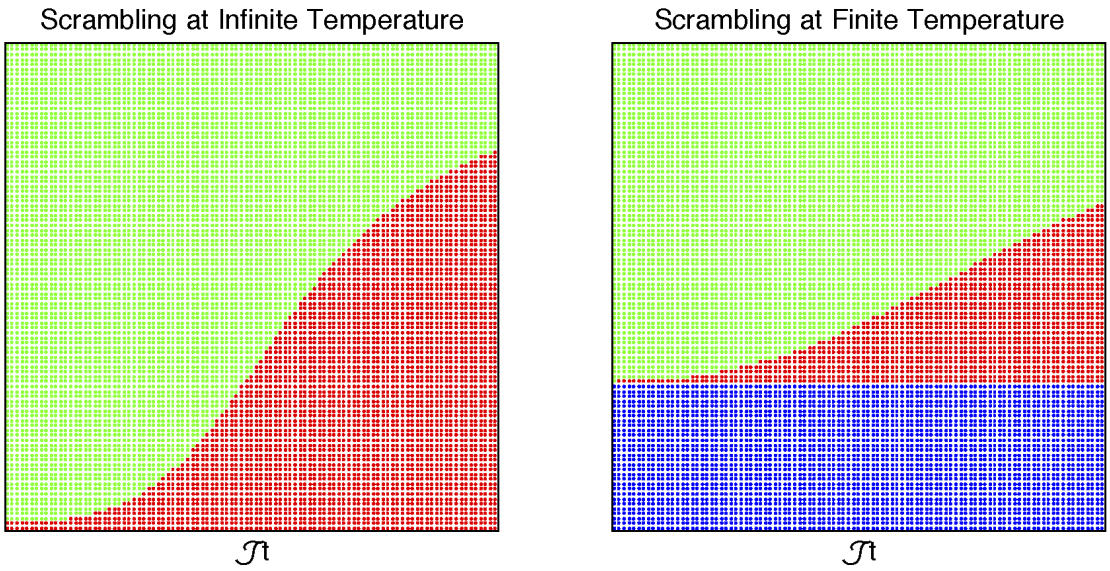


Figure 4.10: Illustration of different epidemics at infinite temperature (left panel) and finite temperature (right panel). Green dots represents unexposed individuals, red dots represents contagious individuals, and blue dots represent vaccinated individuals. Finite temperature factors such as $\rho^{1/2}$ "use up" some of the available flavors for growth, resulting in collisions like those depicted in Fig. (4.5). This effect ends up being well-modeled by an epidemic where these flavors or individuals count as having been exposed, but do not spread disease. As a result of this large vaccinated population, it simply more rare for a contagious individual to encounter an unexposed individual, even at the start. Hence the rate of infection – the Lyapunov exponent – slows down, as seen in the right figure.

consequence that the twisted two-point function solves the ladder kernel [2,14]. In practice solving the former can be substantially easier than solving the latter. As an example, in [113] we use this "twisted" technique to derive an elegant expression for the large-q SYK four-point function at arbitrary coupling and temperature. Like the growth distribution methodology, this new method for calculating four-point functions works for all fermionic systems.

The dynamical renormalization of the coupling (4.50) plays a central role in this work. It will be important to understand this in a deeper and more general

context. The success of the modified infection model in capturing the thermal operator growth suggests that the principle underlying the finite temperature slowdown in SYK is competition for Majorana flavors. The presence of various powers of the thermal state $\exp(-\beta H)$ "uses up" some finite fraction of the flavors. Consequently, when we apply a single fermion, there is a fractional probability for it to become absorbed and thus its size is renormalized (4.33) to a value based upon the percentage δ_β of "unused" flavors. Now, the renormalized coupling $\mathcal{J}(t)$ (4.50) slows down during time-evolution. We believe that this occurs due to the same principle, but have not yet fully understood the mechanism. Our belief is motivated by the empirical observation that the Lyapunov exponent is a power of the percentage δ_β of "unused" flavors

$$\lambda_L = \lim_{t \gg t_{dissipation}} 2\mathcal{J}(t) = 2\mathcal{J}(\delta_\beta)^{q/2} \equiv 2\mathcal{J}\left(1 - \frac{n\left[\rho^{1/2}\right]}{n_*}\right)^{q/2}$$

This kind of sigmoidal operator growth is generic in many-body chaos. However, without Majorana fermions, the manner in which the thermal factors interfere with operator growth must be more complicated, as there is not a bit-like notion of "using up" a flavor. Sigmoidal behavior signals the existence of a competition for some finite resource. For SYK, this resource was flavor; we only have N flavors with which to grow operators, so eventually we will be led to flavor collisions as in Fig. (4.5). However, flavor competition is only one aspect of competition for a more general resource. The question remains: what do operators compete for during evolution? Is there some sort of "operator entropy"? Perhaps when summed across "all operators" at finite temperature, there is always a fixed

amount of total correlation with an initial simple operator due to unitarity. A better understanding of such a resource would give an organization to operator dynamics at different energy scales.

Our results have interesting implications for the holographic dual of the SYK model. This is simplest to understand when we explicitly express our results in terms of the doubled theory. We defined an entangled orthonormal basis for the doubled theory using the eigenstates (4.9) of the size operator (4.13). Taking the state $\psi_1^L(t)|TFD\rangle$[3], we related its size wave-function squared (i.e. $P_n\left[\psi(t)\rho^{1/2}\right] \equiv \left|\langle n|\psi_1^L(t)|TFD\rangle\right|^2$) to the size wave-function squared for the thermofield double state (i.e. $P_n\left[\rho^{1/2}\right] \equiv |\langle m|TFD\rangle|^2$)

$$\left|\langle n|\psi_1^L(t)|TFD\rangle\right|^2 = \sum_{m=0}^{n} K_{n-m}^{\beta}\left[\psi_1(t)\right]|\langle m|TFD\rangle|^2$$

isolating the time-dependence into the growth distribution $K^{\beta}\left[\psi_1(t)\right]$ (4.24). Using this, we found the "average growth" of $\psi_1(t)$ (4.47) at low temperatures to be $1+2\left(\beta\mathcal{J}\sinh\left(\pi t/\beta\right)/\pi\right)^2$, which was shown in [116] to exactly match the classical momentum dynamics of a "boundary" particle falling into a near-extremal black hole. That is, the average growth of an SYK fermion exactly matches the average momentum of an infalling particle in a $NAdS_2$ black hole.

It is a striking result of our analysis that the *full* size wavefunction squared of the SYK fermion precisely relates to the *full* momentum wavefunction squared of the infalling particle. The universal form (4.52) of the growth distribution $K^{\beta}\left[\psi_1(t)\right]$ precisely gives the squared coefficients of the AdS_2 momentum bulk-

[3]If we replace $t \to -t$, then this is the precursor state $\psi_1^L(-t)|TFD\rangle$, where the "boundary" operator ψ_1^L acted upon the thermofield double state at time $-t$ [23, 82, 114, 115].

to-boundary propagator. Exploring this connection will be an important focus of future work[4].

[4]One next step will be to perform this analysis for a different geometry, and a natural place to do so is in the set-up created by [109]. It was found that adding the scrambling distance operator (4.15) to the doubled SYK Hamiltonian causes the low-energy limit to eventually cross a Hawking-Page transition, forming a global $NAdS_2$ geometry instead of a $NAdS_2$ black hole.

Chapter 5

Final Thoughts

With regards to much of the quantum chaos presented in this thesis, there have been developments towards the corresponding bulk constructions. For the spectral form factor $Z(\beta + it) Z(\beta - it)$, a finite ramp - which is the manifestation of eigenvalue repulsion - was obtained in JT gravity by Saad, Shenker, and Stanford through a new semi-classical saddle [117]. As always, the early time behavior is determined by two separate eternal black holes (which naively has four boundaries which are connected into two separate pairs if one wishes to purify the description). However this decays to zero as $t \to \infty$, and eventually there is an additional saddle that takes over. For the spectral form factor $Z(it) Z(-it)$, a saddle point requires two boundaries periodic in *Lorentzian* time. This can be obtained by taking a single two-sided eternal black hole and performing an identification under a finite boost. The resultant "double-cone" geometry no longer has a singularity and produces the late-time linear growth of the spectral form factor. Later, in [118], they found that the plateau, which is a manifestation of the discrete nature of an energy spectra, can also be obtained in gravity through geometrical considerations. Specifically, they found that if one sums over all collections of surfaces containing

two boundaries satisfying the conditions, then one is able to produce a finite constant spectral form factor related to the entropy of the eternal black hole. One should note that these surfaces include those with more than two boundaries, such as those emitting and absorbing "baby universes". In other words, one must sum over all surfaces containing *at least* two boundaries in order to obtain the correct quantum mechanical late time behavior for the spectral form factor.

Furthermore, Lin, Maldacena, and Zhao made progress understanding the behavior of four-point functions and other quantum mechanical observables in the bulk dual of SYK-like models [119]. Specifically, they worked in the quantum mechanical theory of the leading fluctuations around a dominant gravitational saddle: the Schwarzian action. They constructed a set of three gravitationally-dressed operators generating a bulk physical $SL(2,\mathbb{R})$ algebra, moving matter relative to the left and right dynamical boundaries. In the semi-classical limit, these operators agree with those found in [109], with the size operator being the piece of the conformal energy evolving under boundary time translations. Consequently, the agreement between the size distribution of the thermal fermion and the conformal energy distribution of a precursor state seems more sensible. It is worth mentioning that from their construction, the "reason" why various OTOCs saturate at late time is not because there exists a finite number of degrees of freedom, but rather that the two boundaries become purely space-like separated - even at asymptotically late times - due to the infalling particles. In other words, the future null infinities are no longer null-separated from a past null-infinity. As a result, the relative boost angle between the infalling particle and dressed momentum operator is upper bounded.

Alternatively stated, the shockwave due to an infalling particle causes the geodesic distance between the boundary times ($t_L = -t, t_R = t$) to grow, saturating at linear growth with a constant velocity as $t \to \infty$. Thus, the dressed momentum, which is the time derivative of geodesic distance, asymptotes to a constant value. The notion of a distance, as well as a distance operator, for the dominant geometry only seems to work for three and two-dimensional asymptotically AdS geometries. Once in four or higher dimensions, the notion of a minimal geodesic in a black hole background becomes more nuanced, and there does not appear to be an obvious construction in terms of simple boundary operators. It would be interesting to figure out what relation this has, if any, to the existence of bulk gravitons, as well as other phenomena unique to four and higher dimensional gravity. In these cases, it is not possible to perform all the bulk integrals and relate OTOCs to non-trivial null-translated states, unlike the two and three dimensional case. Thus, we expect certain aspects of holography to be qualitatively different in four and higher dimensions due to the existence of bulk gravitational degrees of freedom.

One of the motivations for the operator growth portion of this thesis was to elicit an understanding behind the behavior of four-point functions, which are constrained in factorizing theories to deviate from the disconnected correlator as an exponential which grows no faster than $2\pi/\beta$ in boundary time. From this perspective, it is clear to explain what is occurring. Some initial boundary operator evolves in two ways: it propagates in space, while simultaneously becoming larger products of operators. Once one averages over all possible internal degrees of freedom, the important part of the evolution is the manner in which the bound-

ary operator becomes products of different internal degrees of freedom, at least in all known models of holography. As a result, the operator "scrambles" across the various boundary degrees of freedom.

However, one must note that the average product size of an operator in any subregion is related to the elapsed time; the longer one waits, the bigger an operator becomes. What makes this kind of internal dynamics so different is that it behaves sigmoidally in time, behaving like an epidemic due to the all-to-all nature the interactions. This is similar behavior to what one might for a particle being falling into some horizon, with occupation of the internal degrees of freedom playing the role of an emergent dimension for something to start small and then grow. In CFT language, this can manifest as the behavior of a primary to fall along its descendants in an exponential manner when evolved using one of the non-compact generators.

Of course, the role of the large N degrees of freedom is more complicated than this, but we hope that we have elucidated the manner in which all-to-all interactions naturally create an algebra where things tend to "fall"/grow in size. Essentially, just as spatial derivatives give rise to spatial locality, restricted all-to-all interactions give rise to size/depth/weight locality. It would be interesting to see if it is possible to create such a intuitive perspective for other large N holographic aspects, such as the forming of appropriate conformal gaps, thereby effectively "pulling in" all the black hole hair.

Appendix A

SYK Statistics

A.1 Particle-hole symmetry of SYK

In the Dirac description (2.8) the Hamiltonian has conserved charge parity, where the charge (fermion number) operator is $\widehat{Q} = \sum_i \bar{c}_i c_i$. The Hamiltonian (2.6) has two sectors for charge parity even and odd.

The theory also has a particle-hole symmetry under the operator [35, 36, 43]

$$P = K \prod_{i=1}^{N_d} (\bar{c}_i + c_i) \tag{A.1}$$

where K is the anti-linear operator that takes $z \to \bar{z}$, $z \in \mathbb{C}$ (here we choose $c_i, \bar{c}_i$ to be real with respect to K). One can check that

$$P^2 = (-1)^{\frac{N_d(N_d-1)}{2}} = (-1)^{\lfloor N_d/2 \rfloor} = \begin{cases} +1 & , N_d \bmod 4 = 0 \\ +1 & , N_d \bmod 4 = 1 \\ -1 & , N_d \bmod 4 = 2 \\ -1 & , N_d \bmod 4 = 3 \end{cases} . \tag{A.2}$$

The action on the fermions is given by

$$Pc_iP = \eta\bar{c}_i\,, \quad P\bar{c}_iP = \eta c_i \quad \Rightarrow \quad P\psi_aP = \eta\psi_a\,, \tag{A.3}$$

where

$$\eta = (-1)^{N_d-1}P^2 = (-1)^{\left\lfloor \frac{3N_d}{2}-1\right\rfloor}\,. \tag{A.4}$$

One can now check that P is a symmetry,

$$[H, P] = 0\,. \tag{A.5}$$

For some values of N_d this leads to a degeneracy in the spectrum. P maps a state with fermion number Q to $N_d - Q$ (in our convention the Fock space vacuum has fermion number 0).

1. If $N_d = N/2$ is odd then P maps the even and odd charge parity sectors to each other, and so the two sectors are degenerate.

2. If $N_d = N/2$ is even then P maps each charge parity sector to itself.

   (a) If $(N_d \bmod 4) = 2$ then $P^2 = -1$. Since P is both anti-linear and obeys $P^2 = -1$ it cannot map energy eigenstates states to themselves, and we have double degeneracy within each sector.

   (b) If $(N_d \bmod 4) = 0$ then $P^2 = 1$. In this case there is no protected degeneracy.

Therefore, for $(N \bmod 8) \neq 0$ there is 2-fold degeneracy, while for $(N \bmod 8) = 0$

there is no protected degeneracy.

A.2 The double-scaled SYK theory

In this appendix we compute the disorder-averaged spectrum of the SYK theory in the double-scaled limit

$$N \to \infty, \qquad q \to \infty, \qquad \lambda = \frac{q^2}{N} = \text{fixed}. \tag{A.6}$$

The computation is a small modification of the analysis by Erdős and Schröder in [62] for closely related systems (composed of Pauli matrices with random couplings instead of Majorana operators).

The argument goes as follows: first, we compute the moments $\operatorname{tr} H^k$. Then, we appeal to a combinatoric result in [120] to get the distribution for which these are the moments.

First, we discuss the computation of the moments. We would like to evaluate

$$\langle \operatorname{tr} H^k \rangle_J \tag{A.7}$$

for k even. We evaluate the J integral by Wick contractions. This involves pairing up the various terms in different H factors and contracting the flavor indices of the fermions that appear in the pair. If all of the Wick-contracted pairs were adjacent in the product, we could evaluate each pair as $\frac{1}{2^q}$, since $\psi_i \psi_i = \frac{1}{2}$. Taking the product over the pairs and summing over the possible fermion flavors that can

occur in each pair, we get

$$\frac{\operatorname{tr} H^k_{\text{assuming all pairs next to each other}}}{\operatorname{tr} 1} = \left[\frac{\langle J^2\rangle}{2^q}\binom{N}{q}\right]^{k/2} = \left(\frac{\mathcal{J}^2}{2\lambda e^{\lambda/2}}\right)^{k/2}. \tag{A.8}$$

where $\mathcal{J}$ is defined by $\langle J^2_{i_1\ldots i_q}\rangle = \frac{2^{q-1}}{q}\frac{\mathcal{J}^2(q-1)!}{N^{q-2}}$ [2]. Now, of course we also have to consider cases where Wick-contracted pairs are not adjacent. The procedure is to commute the terms past each other until the contracted pairs are adjacent or nested, so that Wick-contraction lines do not cross.

Let's consider what happens when we move one product of fermions past another. Notice that

$$\left[\psi_{a_1}\ldots\psi_{a_q}\right]\left[\psi_{b_1}\ldots\psi_{b_q}\right] = (-1)^{\#\text{ fermions in common}}\left[\psi_{b_1}\ldots\psi_{b_q}\right]\left[\psi_{a_1}\ldots\psi_{a_q}\right]. \tag{A.9}$$

The important feature of the limit where we hold q^2/N fixed is that the expected number of fermions in common stays of order one in this limit. More precisely, the number is Poisson distributed, with distribution

$$P(m \text{ fermions in common}) = \frac{\lambda^m}{m!}e^{-\lambda}, \qquad \lambda = \frac{q^2}{N}. \tag{A.10}$$

Now, in principle, things will get complicated because we have to consider the possibility that the same fermions that are shared between two copies of the Hamiltonian containing $\psi_{a_1}\cdots\psi_{a_q}$ and $\psi_{b_1}\cdots\psi_{b_q}$ might also be shared with a third copy containing $\psi_{c_1}\cdots\psi_{c_q}$. Or, more generally, that the number of such terms might be correlated. However, the probability is proportional to $1/N$, without a q^2 enhancement, so we ignore it in the double-scaled limit. This is the key point

that makes it possible to solve.

So now, each time we have to commute a product of q fermions past each other, we get a factor

$$\sum_{m=0}^{\infty}(-1)^m P(m \text{ fermions in common}) = e^{-2\lambda}. \tag{A.11}$$

Notice that this step differs somewhat from the case considered by [62]. Specifically, this is where the fact that we have Majoranas instead of spins is relevant. In the spin case, the analogous sum gives $e^{-4\lambda/3}$. Anyhow, doing this sum independently for each set of fermions that we need to commute past each other, we can now correct the expression (A.8), and we find

$$\frac{\operatorname{tr} H^k}{\operatorname{tr} 1} = \sum_{\text{Wick pairings}} \left(\frac{\mathcal{J}^2}{2\lambda e^{\lambda/2}}\right)^{k/2} e^{-2\lambda \text{ cross(pairing)}}. \tag{A.12}$$

Here cross() gives the number of commutations required to get the pairs arranged in a way so that they are all adjacent or nested. We can describe this target situation by saying that lines connecting the Wick pairs will not cross. Then cross() is just the initial number of crossings of Wick contraction lines.

The final step is to notice [62] that the distribution with these moments is known [120]. It is related to the integration measure for the Q-Hermite polynomials, with $Q = e^{-\lambda}$. The distribution is given by

$$\rho(E) = \frac{\mathcal{N}}{\sqrt{1-a^2}} \prod_{n=0}^{\infty}\left(1 - \frac{a^2}{\cosh^2(n\lambda)}\right), \qquad a^2 \equiv \frac{\lambda e^{\lambda/2}(1-e^{-2\lambda})}{2}\frac{E^2}{\mathcal{J}^2}. \tag{A.13}$$

for $|a| < 1$ and zero otherwise. The normalization factor can be determined from

the constraint that the total number of states is $2^{N/2}$.

It is convenient to rewrite ρ as follows

$$\log \frac{\rho(E)}{\mathcal{N}} = \frac{1}{2} \sum_{n=-\infty}^{\infty} \log \left(1 - \frac{a^2}{\cosh^2(n\lambda)}\right) \tag{A.14}$$

$$= \frac{1}{2} \sum_{k=-\infty}^{\infty} \int_{-\infty}^{\infty} dn\, e^{-2\pi i k n} \log \left(1 - \frac{a^2}{\cosh^2(n\lambda)}\right) \tag{A.15}$$

$$= -\frac{1}{\lambda} (\arcsin a)^2 + \sum_{k\geq 1} \frac{1 - \cosh\left[\frac{k\pi}{\lambda}(\pi - 2\arccos a)\right]}{k \sinh \frac{k\pi^2}{\lambda}}. \tag{A.16}$$

In the second line we used the Poisson resummation formula. In the last line we did the n integral by contour integration, summing over a geometric series of cuts of finite length along the imaginary n axis. This formula is now in a convenient form for discussing the behavior at small λ.

For example, if we take $\lambda \to 0$ with a fixed, the first term dominates, and exactly reproduces the large q thermodynamics computed in [2].

Our primary goal is to use this to evaluate the partition function of the Schwarzian theory, so we take a further "triple-scaled" limit

$$\lambda \to 0, \qquad \frac{E - E_0}{\mathcal{J}} \to 0, \qquad z \equiv \frac{(E - E_0)}{\lambda \mathcal{J}} = \text{fixed}. \tag{A.17}$$

In this limit, we approximate $a = -1 + z\lambda^2 + O(\lambda^3)$ and we have

$$\frac{\arccos(a)}{\lambda} = \sqrt{2z} + O(\lambda), \qquad \frac{(\arcsin(a))^2}{\lambda} = \frac{\pi^2}{4\lambda} - \pi\sqrt{2z} + O(\lambda)\,, \quad \text{(A.18)}$$

$$\sum_{k\geq 1} \frac{1 - \cosh\left[\frac{k\pi}{\lambda}\left(\pi - 2\arccos a\right)\right]}{k \sinh\frac{k\pi^2}{\lambda}} = \log\left(1 - e^{-2\pi\sqrt{2z}+O(\lambda)}\right) + O(e^{-\pi^2/\lambda}). \quad \text{(A.19)}$$

We conclude that in the triple-scaled limit we have

$$\rho(E) = 2\mathcal{N} e^{-\frac{\pi^2}{4\lambda}} \sinh\left(\pi\sqrt{2z}\right), \qquad z = \frac{(E - E_0)}{\lambda\mathcal{J}}, \qquad \lambda = \frac{q^2}{N}. \quad \text{(A.20)}$$

One can check that for small λ the normalization factor is $\mathcal{N} = \frac{2^{N/2}}{\mathcal{J}}\sqrt{\frac{\lambda}{\pi}}$, which leads to

$$Z(\beta) = \int dE \rho(E) e^{-\beta E} = e^{-\beta E_0 + S_0} \frac{\sqrt{2}\pi}{(\beta\mathcal{J})^{3/2}} \exp\left(\frac{\pi^2}{2\lambda\beta\mathcal{J}}\right). \quad \text{(A.21)}$$

where $E_0 = -\frac{\mathcal{J}}{\lambda}$ and $S_0 = N\frac{\log(2)}{2} - \frac{\pi^2}{4\lambda}$. This agrees with the 1-loop calculation of [2], but here we conclude that it is the exact answer in the triple-scaled limit that isolates the Schwarzian.

Finally, we will mention that there is another way to analyze the double-scaled limit, starting from the G, Σ action for the disorder-averaged partition function:

$$-I = \frac{N}{2}\log\det(\partial_\tau - \Sigma) - \frac{N}{2}\int d\tau_1 d\tau_2 \left[\Sigma G - \frac{\mathcal{J}^2}{2q^2}(2G)^q\right]. \quad \text{(A.22)}$$

To take the double-scaled limit, we write

$$\Sigma(\tau_1, \tau_2) = \frac{\sigma(\tau_1, \tau_2)}{q}, \qquad G(\tau_1, \tau_2) = \frac{\text{sgn}(\tau_{12})}{2}\left(1 + \frac{g(\tau_1, \tau_2)}{q}\right). \tag{A.23}$$

where now $g(\tau_1, \tau_2)$ is a symmetric function of its two arguments that is constrained to vanish when they coincide. The action in the double-scaled limit is

$$-I = \frac{N}{4q^2}\Big[- \int d\tau_1..d\tau_4 \frac{\text{sgn}(\tau_{12})}{2}\sigma(\tau_2, \tau_3)\frac{\text{sgn}(\tau_{34})}{2}\sigma(\tau_4, \tau_1) \tag{A.24}$$

$$+ \int d\tau_1 d\tau_2 \left(\mathcal{J}^2 e^{g(\tau_1,\tau_2)} - \text{sgn}(\tau_{12})\sigma(\tau_1, \tau_2)g(\tau_1, \tau_2)\right)\Big]. \tag{A.25}$$

Notice that σ appears quadratically, so we can integrate it out exactly. We get

$$-I = \frac{N}{4q^2}\int d\tau_1 d\tau_2 \left[\mathcal{J}^2 e^{g(\tau_1,\tau_2)} - \frac{1}{4}\partial_{\tau_1} g(\tau_1, \tau_2)\partial_{\tau_2} g(\tau_1, \tau_2)\right], \tag{A.26}$$

which has the form of a Liouville action on a Lorentzian space. One can analyze this theory by studying perturbation theory in $\mathcal{J}^2$. This is equivalent to computing moments as in the Erdos-Schroder analysis. Note that at each order in $\mathcal{J}^2$ we have a simple Gaussian integral.

A.3 A toy G, Σ integral

In the main text, we asserted that G, Σ give a nonperturbatively exact formulation of the disorder-averaged SYK model. In this appendix, we discuss a toy model for the G, Σ path integral. We discuss the contour of integration and saddle points, and we see how Grassmann behavior can arise from bosonic variables.

The example that we will discuss can be thought of as the SYK Grassmann path integral on a space where we replace the time dimension by two points, labeled 1 and 2. Then the fermion variables are $\psi_i(1), \psi_i(2)$ where $i = 1, ..., N$. Concretely, the integral we consider for fixed disorder is

$$Z = \int d^N\psi(1) d^N\psi(2) e^{\sum_i \psi_i(1)\psi_i(2) + \sum_{i_1<...<i_q} J_{i_1...i_q}[\psi_{i_1}(1)...\psi_{i_q}(1) + \psi_{i_1}(2)...\psi_{i_q}(2)]}. \tag{A.27}$$

The average over couplings gives

$$\langle Z \rangle = \int d^N\psi(1) d^N\psi(2) e^{\sum_i \psi_i(1)\psi_i(2) + \frac{(q-1)!J^2}{N^{q-1}} \sum_{i_1<...<i_q} \psi_{i_1}(1)\psi_{i_1}(2)...\psi_{i_q}(1)\psi_{i_q}(2)} \tag{A.28}$$

$$= \int d^N\psi(1) d^N\psi(2) e^{\sum_i \psi_i(1)\psi_i(2) + \frac{J^2}{qN^{q-1}}[\sum_i \psi_i(1)\psi_i(2)]^q}. \tag{A.29}$$

We can write this as a G, Σ integral by the standard manipulation: we introduce a variable σ that is a Lagrange multiplier that sets $g = \frac{1}{N}\sum_i \psi_i(1)\psi_i(2)$. This leads to the expression

$$\langle Z \rangle = N \int d^N\psi(1) d^N\psi(2) e^{\sum_i \psi_i(1)\psi_i(2)} \int dg \frac{d\sigma}{2\pi i} e^{\sigma[\sum_i \psi_i(1)\psi_i(2) - Ng] + \frac{J^2}{q} g^q} \tag{A.30}$$

$$= N \int dg \frac{d\sigma}{2\pi i} e^{N[\log(1+\sigma) - \sigma g + \frac{J^2}{q} g^q]}. \tag{A.31}$$

We would now like to describe how to make sense of this integral. The defining contour of integration for σ is along the imaginary axis, and we start by formally integrating g along th real axis. We then evaluate the integral as follows: if we bring down the $\log(1 + \sigma)$ term and expand in powers of σ, we will have σ integrals

of the form $\frac{N}{2\pi i}\int d\sigma \sigma^p e^{-N\sigma g} = N^{-p}(-\partial_g)^p \delta(g)$. This leads to

$$\langle Z\rangle = \int dg e^{N\frac{J^2}{q}g^q}(1-N^{-1}\partial_g)^N \delta(g) = (1+N^{-1}\partial_g)^N e^{N\frac{J^2}{q}g^q}\Big|_{g=0} \tag{A.32}$$

$$= \sum_{m=0}^{\lfloor N/q\rfloor} \frac{N!}{(N-mq)!m!}\left(\frac{J^2}{N^{q-1}q}\right)^m. \tag{A.33}$$

This is the right answer, and we got it from an integral over bosonic variables, but the final g integral was supported in a neighborhood of the origin, and the calculation reduced rather trivially to a direct fermionic computation of (A.29).

However, we can also change the contour and make the integral more manifestly well-defined. We rotate the g and σ contours in opposite directions by $e^{i\pi/q}$. Here it is simplest to define new variables $\tilde{\sigma} = -i\sigma e^{-i\pi/q}$ and $\tilde{g} = e^{i\pi/q}g$. Then we have

$$\langle Z\rangle = N\int d\tilde{g}\frac{d\tilde{\sigma}}{2\pi} e^{N[\log\left(1+ie^{i\pi/q}\tilde{\sigma}\right)-i\tilde{\sigma}\tilde{g}-\frac{J^2}{q}\tilde{g}^q]}, \tag{A.34}$$

where we integrate $\tilde{g}, \tilde{\sigma}$ over the real axis. It is easy to check that numerical integration first over $\tilde{g}$ and then over $\tilde{\sigma}$ indeed gives the correct answer (A.33) for a few values of N, q.

One can also discuss saddle points for this integral. For these purposes we go back to the g, σ variables. There are q saddle points, the solutions of the equations

$$\sigma = J^2 g^{q-1}, \qquad g = \frac{1}{1+\sigma}. \tag{A.35}$$

There is one real solution, and this is the one that naively dominates. We have not analyzed the deformation of the integration contour in detail, but we observe that this leading saddle does in fact give the right large N behavior, comparing

to (A.33).

A confusing aspect of the G, Σ representation is that the fundamental variables are Grassmann variables, and we could ask how this is consistent with a representation by g, σ. For example, the fact that the square of a Grassmann vanishes should imply that $g^{N+1} = 0$. This seems inconsistent with the fact that we are integrating over nonzero values of $\tilde{g}$, and indeed studying saddle points with g nonvanishing. In fact, one can check that an insertion of g^p with $p > N$ will make the integral zero. This is easiest to see from (A.32), based on the fact that we are at most taking N derivatives of the integrand before setting $g = 0$, so a term of degree $N + 1$ will give zero.

A.4 Subleading saddle points in the G, Σ variables

Besides the standard saddle point that gives the themodynamics discussed in section 2.4, there are a family of subleading saddles for the path integral (2.16). We do not have their explicit form for $q = 4$, but we can understand some of their properties numerically, and by comparison to the simpler $q = 2$ theory.

In the $q = 2$ model, the saddle point equations for different Matsubara frequencies decouple, and we have

$$G(\omega_n)^{-1} = -i\omega_n - \Sigma(\omega_n), \qquad \Sigma(\omega_n) = J^2 G(\omega_n) \tag{A.36}$$

with solutions

$$G_{\pm}(\omega_n) = \frac{-i\omega_n \pm i\text{sgn}(\omega_n)\sqrt{\omega_n^2 + 4J^2}}{2J^2}. \tag{A.37}$$

Choosing G_+ for all frequencies gives the dominant saddle. Choosing G_- for some of the frequencies will lead to subdominant saddles. The difference in saddle point action induced by choosing G_- (for both $\omega_n = \frac{2\pi}{\beta}(n + 1/2) > 0$ and the corresponding $-\omega_n$) is

$$-I(G_+) + I(G_-) = N \log \frac{1 + \sqrt{4J^2/\omega_n^2 + 1}}{1 - \sqrt{4J^2/\omega_n^2 + 1}} + N\frac{|\omega_n|\sqrt{4J^2 + \omega_n^2}}{2J^2} \tag{A.38}$$

$$= N\left(i\pi + \frac{4\pi(n + \frac{1}{2})}{\beta J} + O(\frac{1}{(\beta J)^3})\right). \tag{A.39}$$

For large βJ, we see that the saddles become almost degenerate. Naively, this would suggest a soft mode connecting the saddles, but because the imaginary part differs by an order one amount, we do not have such a mode. However, at large $\beta J \gg N$ one would have to sum over all of these saddles. We will see that they play an important role in appendix A.5.

In the $q = 4$ model we do not have explicit formulas, but we can find subleading saddles numerically. It seems that for each $q = 2$ solution there is a corresponding $q = 4$ solution, which can be found by starting with the $q = 2$ solution and iterating the Schwinger-Dyson equations while slowly increasing q from two to four. We give a plot of some solutions in Fig. A.1. An important difference between the $q = 2$ and $q = 4$ cases is that the actions do not become degenerate at large βJ. For the simplest case, where we start with a $q = 2$ solution with a single frequency pair ω_n flipped, we find numerically that the $q = 4$ action is given

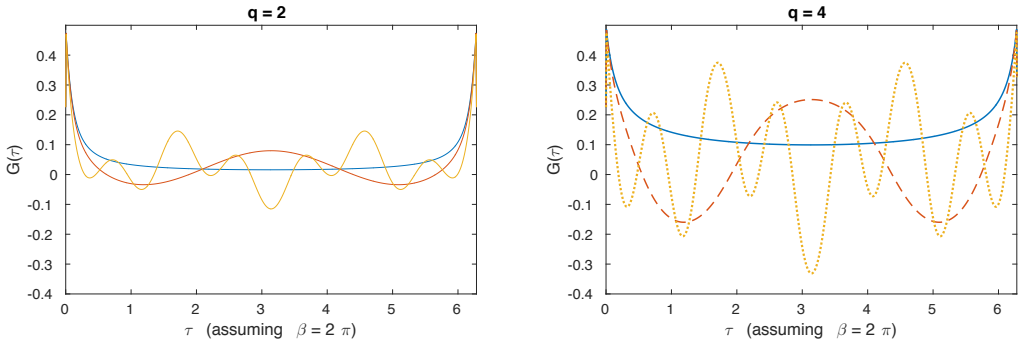


Figure A.1: At left we show the $q = 2$ standard saddle (blue/solid) and a subleading saddle with $n = 1$ flipped (red/dashed) and one with both $n = 2, 6$ flipped (orange/dotted). At right we have the corresponding $q = 4$ solutions. We use $\beta\mathcal{J} = 20\pi$.

by

$$-I(G_{standard}) + I(G_{subleading}) \approx N\left(i\pi + \frac{n+1}{2} + O(\frac{1}{\beta J})\right). \tag{A.40}$$

We are not sure that this simple expression is exactly correct, only that it is within a percent or so of the numerical answer for the first few frequencies $n = 0, ..., 5$ where we were able to check. The important point is that there is a large $\frac{N}{2}$ gap in the action even at very low temperature. This explains why these additional saddles do not disturb the large N thermodynamics. A logical possibility is that the relative dominance of these saddles could change when we study complex β, but preliminary investigation suggests that this is not the case, and that the gap remains.

Finally, we will mention that in the $q = 4$ theory there also appear to be saddle points that depend nontrivially on both of the time arguments, not just the difference. In other words, we have saddle points that spontaneously break time translation invariance. We have not studied these systematically, but the

examples that we found numerically had larger action than the standard saddle.

A.5 Saddle points and the $q = 2$ model

The $q = 2$ model is qualitatively different than the model with $q > 2$, since it is equivalent to a model of free fermions with a random mass matrix [30, 121, 122]. It is not a chaotic system, but the explicit $N \times N$ random matrix leads to a "mini-ramp" and "mini-plateau" in certain quantities, with plateau time $t_p \sim N$ instead of $t_p \sim L$. In this appendix we show how the saddle points discussed in Appendix A.4 contribute to this behavior.

The Hamiltonian of the $q = 2$ model is

$$H = i \sum_{i<j} J_{ij} \psi_i \psi_j, \tag{A.41}$$

where J_{ij} is a real antisymmetric matrix. Conjugating with an orthogonal matrix Q, we can take J_{ij} to a block diagonal form with each block given by

$$\begin{pmatrix} 0 & \lambda_k \\ -\lambda_k & 0 \end{pmatrix} \tag{A.42}$$

where $\lambda_k > 0$ and with k running from 1 to $N/2$. Then the Hamiltonian can be written as

$$H = i \sum_{k=1}^{N/2} \lambda_k \tilde{\psi}_{2k-1} \tilde{\psi}_{2k} = \sum_{k=1}^{N/2} \lambda_k (c_k^\dagger c_k - \frac{1}{2}). \tag{A.43}$$

Where $\tilde{\psi}_i = (Q\psi)_i$, and we made Dirac fermions out of these pairs of Majoranas, $c_k = (\tilde{\psi}_{2k-1} + i\tilde{\psi}_{2k})/\sqrt{2}$ and $c_k^\dagger = (\tilde{\psi}_{2k-1} - i\tilde{\psi}_{2k})/\sqrt{2}$. Notice that iJ_{ij} is a skew

Hermitian matrix, not a GUE Hermitian matrix. Its eigenvalue statistics are known [30, 40]. At large N the spectrum is a semicircle, with $1/N$ corrections. Eigenvalue (mass) pair correlations $R_2(\lambda, \lambda')$ are described by a modified sine kernel whose short distance behavior is that of GUE.

It follows that eigenvalues in the single particle sector will repel, because of the usual eigenvalue repulsion of a random matrix. However, nearby multiparticle eigenvalues coming from sectors with very different particle numbers will repel only weakly. Because the eigenvalues that repel each other have an average spacing $\sim 1/N$ instead of $1/L$, we expect that the plateau time in this model is $t_p \sim N$.

The simplest observable in this model with a ramp is the (quenched) disorder averaged squared correlation function. It turns out this is easier to calculate than $g_d(t)$. The averaged correlation function (not squared) does not have a ramp.

Part of the reason that the correlation functions are easier to calculate is that the matrix elements of ψ_i, $\langle n|\psi_i|m\rangle$, are only nonzero for $|n\rangle$, $|m\rangle$ belonging to particle number sectors differing by a particle number of one. This means that the correlation function only receives contributions from energy differences that are equal to the single particle sector energies, making the calculation much simpler. Explicitly, the Euclidean quenched correlation function is

$$G(\tau) = \frac{1}{N}\sum_{i=1}^{N}\left\langle \frac{\mathrm{Tr}[e^{-\beta H}\psi_i(\tau)\psi_i]}{Z(\beta)}\right\rangle_J = \int d\lambda \frac{1}{e^{-\beta\lambda}+1}\tilde{\rho}(\lambda)e^{-\lambda\tau} \tag{A.44}$$

Here $\tilde{\rho}(\lambda)$ is the average mass density. In the above integral we are extending it to a symmetric function $\tilde{\rho}(-\lambda) = \tilde{\rho}(\lambda)$. We can see that the real time correlation

function, obtained by continuing $\tau \to it$ in the above expression, will not have a ramp or plateau. However, the connected part of the quenched disorder averaged square of the correlation function will have a ramp and plateau

$$G_c^2(\tau, \tau') = \frac{1}{N^2} \sum_{i,j=1}^{N} \left\langle \frac{\mathrm{Tr}[e^{-\beta H}\psi_i(\tau)\psi_i]\mathrm{Tr}[e^{-\beta H}\psi_j(-\tau')\psi_j]}{Z(\beta)^2} \right\rangle_J - G(\tau)G(-\tau') \tag{A.45}$$

$$= \int d\lambda d\lambda' \frac{1}{e^{-\beta\lambda}+1} \frac{1}{e^{-\beta\lambda}+1} R_2(\lambda, \lambda') e^{-\lambda_1\tau + \lambda_2\tau'} \tag{A.46}$$

Note that the annealed correlator cannot be written simply in terms of $R_2(\lambda, \lambda')$. After analytically continuing $\tau \to it$ and $\tau' \to it'$, because of the presence of $R_2(\lambda, \lambda)'$, $G_c^2(t, t')$ will have a ramp and plateau. In particular, at $\beta = 0$ it is precisely equal to $g_c(t)$ for the ensemble of skew Hermitian matrices.

This simple expression for the square of the averaged correlation function in terms of the mass pair correlator suggests that it may be possible to calculate in a simple way by saddle point. Kamenev and Mezard [60] calculated $R_2(\lambda, \lambda')$ in the GUE by saddle point with an integral that is very similar to the path integral (2.15) with $q = 2$.[1] This is why we want to consider the quenched disorder averaged correlation function instead of the annealed disorder average correlation function (where we would J average the denominator and numerator in (A.45) separately).

The quenched disorder averaged squared correlation function in Matsubara frequency space, $G_c^2(\omega_n, \omega_m)$, can be calculated by coupling sources z_n to the

[1] The integral that [60] calculated is closer to the integral (2.15) over only one of the Matsubara frequency modes

operators $\sum_{i=1}^{N} \psi_i(-\omega_n)\psi_i(\omega_n)$ with a term in the action

$$S \supset \frac{1}{2}\sum_{i=1}^{N}\sum_{n=0}^{\infty} \psi_i(-\omega_n)\psi_i(\omega_n)z_n.$$

Let $Z(\{z\})$ be the partition function with the source term included,

$$G_c^2(\omega_n, \omega_m) = \tag{A.47}$$

$$\frac{1}{N^2}\frac{\partial}{\partial z_n}\frac{\partial}{\partial \tilde{z}_m}\left(\left\langle \log Z(\{z\}) \log Z(\{\tilde{z}\})\right\rangle_J - \left\langle \log Z(\{z\})\right\rangle_J \left\langle \log Z(\{\tilde{z}\})\right\rangle_J\right)\Big|_{\{z\},\{\tilde{z}\}=0} \tag{A.48}$$

The key simplification is that the partition function is a product over all the frequencies, $Z(\{z\}) = \prod_{n=0}^{\infty} Z_n(\{z_n\})$, and since the logarithms of these products turn into sums over the different frequencies, the derivatives simplify. We find

$$G_c^2(\omega_n, \omega_m) = \tag{A.49}$$

$$\frac{1}{N^2}\frac{\partial}{\partial z_n}\frac{\partial}{\partial \tilde{z}_m}\left(\left\langle \log Z_n(z_n) \log Z_m(\tilde{z}_m)\right\rangle_J - \left\langle \log Z_n(z_n)\right\rangle_J \left\langle \log Z_m(\tilde{z}_m)\right\rangle_J\right)\Big|_{z_n,\tilde{z}_m=0} \tag{A.50}$$

Now we evaluate the averaged logarithms of the single frequency factors of the sourced partition function. This is almost exactly the calculation of Kamenev and Mezard [60].[2] They use the replica trick to rewrite the average of the logarithm in terms of the average of the replicated partition function. They then evaluate the replicated partition function by the saddle point approximation. Their saddle point equation for the average of a single logarithm is equivalent to the equation

[2]Their calculation applied to the GUE, while ours applies to the ensemble of skew Hermitian matrices. The difference between our integrals comes from the reality constraint on the Majoranas, which gives a different result at order 1/N.

obtained by combining the saddle point equations for $G(\omega_n)$ and $\Sigma(\omega_n)$ for one frequency (A.36), except that we now account for the source with a shift of ω_n, $-i\omega_n \to -i\omega_n + z_n$.

For the average of the product of logarithms, the saddle point equations have a mixing term. These equations are quadratic and thus have two solutions, $G_+(\omega_n)$ and $G_-(\omega_n)$ (A.37). Choosing a replica symmetric solution with G_+ for each replica gives the dominant contribution to the integrals, the fluctuation is the first term that survives. These contributions correspond to the semicircle part of the mass distribution and mass pair correlation function, and the fluctuations give the ramp. Considering a replica symmetry breaking saddle point involving both $G_+(\omega_n)$ and $G_-(\omega_n)$ gives the sine kernel type contribution to $R_2(\lambda, \lambda')$ in $G_c^2(\omega_n, \omega_m)$, and thus gives us the plateau.

As we noted above, calculations of $Z(t)$ and $g(t)$ using saddle points will be more complicated. It appears that the Itzykson-Zuber integral [123] will be helpful. We hope to return to this issue in future work.

A.6 On N^{-q} vs. 2^{-N}

It would be nice to have a direct analytical argument for the ramp and plateau in SYK. As a first step, one would like to understand where the e^{-2S} scale of the ramp comes from. Naively, this is puzzling, because the ramp arises from correlations between the two replicas, and in simple diagrams such correlations are suppressed by powers of N^q, not exponential factors. In this appendix, we make a simple comment about how the exponential can emerge from such diagrams.

We start by defining the quantity

$$F_{k_1,k_2} \equiv \frac{\langle \operatorname{tr} H^{k_1} \operatorname{tr} H^{k_2} \rangle}{L^2 \sigma^{k1+k2}}, \qquad \sigma^2 = \frac{1}{L}\langle \operatorname{tr} H^2 \rangle. \tag{A.51}$$

In principle, knowing F_{k_1,k_2} makes it possible to evaluate the double resolvent $\langle \operatorname{tr} \frac{1}{z-H} \operatorname{tr} \frac{1}{w-H} \rangle$. By taking discontinuities in both z and w across the real axis, one gets an expression for the pair correlation function $\langle \rho(z)\rho(w) \rangle$, which gives rise to the ramp.

This procedure has been carried out for the GUE ensemble by Brezin and Zee [58]. At leading order in $1/L^2$, one considers planar graphs only: most of the Wick contractions do not contribute, and many of the remaining graphs for the double resolvent can be summed by replacing z, w by dressed propagators. All that remains is a special class of graphs where we take $k_1 = k_2 = k$ and then Wick-pair the Hamiltonians in (A.51) "straight across" up to an overall reflection. More explicitly, the first H factor in the first trace is paired with the k-th factor in the second trace. The second factor in the first trace is paired with the $(k-1)$-st factor in the second trace, and so on. We refer to the result of this special contraction as f_k.[3] In GUE one finds $f_k = 2^{-N}$, which is the origin of the 2^{-N} coefficient of the ramp. The linear time dependence arises from a singularity in the geometric series that defines the double resolvent, and in particular is sensitive only to the f_k for large k. This is an important point so we will emphasize it: the short-distance correlations between eigenvalues, or equivalently the late-time behavior of the ramp, is related to the large k behavior of the f_k or F_{k_1,k_2} coefficients.

[3] The contribution of all contractions related to such a configuration by cyclicity would be kf_k.

In SYK, the class of graphs that must be summed at leading order is larger than in GUE. In particular, we have to think about the $1/N$ expansion instead of the $1/L$ expansion. We will not attempt to analyze the sum in a systematic way. Instead, we will simply comment on the behavior of the special class of graphs that we used to define f_k above, because these already provide a model for the handoff between N^{-q} and 2^{-N}.

We define f_k the same way as above: we let $k_1 = k_2$ in (A.51) and we Wick-contract the couplings in each factor of H in the first trace with the corrresponding (reflected, as before) factor of H in the second trace. This is equivalent to the following: we imagine writing a product of k of the possible terms that appear in the Hamiltonian. Then f_k is simply the probability that such a product has a nonzero trace. For small values of k, f_k is suppressed by powers of N^q, as expected for a two-replica correlation. For example, $f_2 = \frac{1}{\binom{N}{q}} \sim N^{-q}$. However, for large values of k, f_k approaches a constant value of 2^{1-N}. This is because for a product of fermions to have a nonzero trace, we must have an even number of each flavor of fermion, leading to N binary constraints. The exact formula is

$$\begin{aligned} f_k &= 2^{-N} \sum_{m=0}^{N} \binom{N}{m} \left(\sum_{p=0}^{q} \binom{m}{p} \binom{N-m}{q-p} (-1)^p \right)^k \\ &= 2^{-N} \sum_{m=0}^{N} \binom{N}{m} \alpha^k, \quad \alpha \equiv \frac{(N-q)!(N-m)!}{N!} \frac{{}_2F_1(-m,-q,N-m-q+1,-1)}{\Gamma(N-m-q+1)}. \end{aligned}$$

For large values of k the largest α dominates the sum. This is the value $\alpha = 1$ when $m = 0, N$, which leads to $f_k \approx 2^{1-N}$. In particular, for the large values of k relevant for the late-time ramp, we find the same behavior as in GUE, for any value of q. We suspect that this is a hint of the universality of local random

matrix statistics, and is the basic point behind the origin of the $e^{-2S} \sim 2^{-N}$ ramp.

We are currently working to make this more precise.

A.7 Constraints on saddle point origins of the ramp

As explained in Section 2.5, the late time plateau is a highly non-perturbative effect in SYK that is expected to involve effects as small as $\exp\left(-e^{N}\right)$, based on random matrix theory analysis. On the other hand, the ramp scales as e^{-N} and so it may be a more tractable non-perturbative effect. In particular, random matrix theory tells us that the part of the ramp that is linear in time is a perturbative effect in RMT, and this part may be an ordinary non-perturbative effect in SYK.

In this appendix we make a few comments about the simplest possible approach to explaining the ramp – finding a nontrivial saddle of the original G, Σ action. But because G is small the source $\log G$ in the action will deform the saddle point. There is backreaction.

Such a saddle would have to satisfy constraints. First, in order to account for the N mod 8 periodicity discussed in Section 2.7 there would have to multiple saddles with complex action.

The second constraint is more nontrivial. As discussed in Section 2.8, the ramp and plateau are not self-averaging (both in the two-point function and in the spectral form factor) [69]. The fluctuations on the ramp are of the same size as its mean value. But a saddle point explanation requires that we have a limit in which fluctuations are suppressed.

This argument may seem a bit quick because the large fluctuations we are discussing are in the integral over random couplings, but this integral can be performed exactly. In particular, in the G, Σ formulation the disorder integral is done first, followed by the integral over the fermion variables, and we are left with an integral over the G, Σ variables. We checked that the latter integral also exhibits large fluctuations on the ramp and plateau (of the same order as the mean value), by numerically comparing the variance $\langle G(t)^2 \rangle - \langle G(t) \rangle^2$ with the mean, directly in the original fermion formulation.

It is possible that the saddle point backreaction for $\langle G \rangle$ and for $\langle G^2 \rangle$ is delicately tuned to make these answers consistent with numerics, but we see no obvious mechanism for this.

A.8 Data

This section contains some further numerical results. We first present $g(t)$, $g_c(t)$, and $g_d(t)$ for $\beta = 0, 1, 5$ for $N = 16, 18, \ldots, 34$ and discuss the dip-ramp-plateau features of g and g_c, which exhibit the mod-8 symmetry pattern. The methods for determining the dip time t_d and the plateau time t_p are explained next, with the results for $N = 10, 12, \ldots, 34$. We compare the fit of t_d with an exponential and a power-law function. The error bars are large but the results for larger N are consistent with the estimate in Section 2.6. They are also consistent with other scenarios involving a crash at earlier time. The available N values are not large enough to disentangle all these effects.

The plateau time t_p shows a faster exponential increase, and the numerical

result is compared with the results of Sections 2.4 and 2.6. This, together with the results for t_d, show that the ramp length grows exponentially in N. For $N = 34$ we have fitted the ramp power law omitting times near t_p where unfolding effects are important. We find a power consistent with the GUE behavior $g(t) \sim t^1$ within a couple percent.

All $g(t)$, $g_c(t)$, and $g_d(t)$ data discussed so far has been for factorized (annealed) quantities, as in (2.10)–(2.12). We compare with the results of the unfactorized (quenched) versions in Sec. A.8.3.

Finally, in A.8.4 we plot the average density of states for different values of N.

A.8.1 Plots of $g(t)$, $g_c(t)$, and $g_d(t)$

In Fig. A.2 we plot $g(t)$, $g_c(t)$, and $g_d(t)$ on a log-log scale. The oscillation observed for $\beta = 0$ before the dip time is also visible for $\beta = 1$ but becomes negligible for $\beta = 5$. It is due to interference between the upper and lower edges of the eigenvalue distribution.

$g_d(t)$ decays quickly to typically much smaller values than $g(t)$ or $g_c(t)$ around the dip time. This is consistent with the theoretical expectation of a gaussian falloff due to fluctuations in the edge of the eigenvalue distribution at times of order N (albeit with a somewhat large coefficient). Such effects cancel out in $g(t)$. (Beyond the dip time $g_d(t)$ seems to rebound. This is just because the number of samples is finite and hence the cancellation is not perfect.)

Around the plateau time, the curves for $g(t)$ and $g_c(t)$ exhibits a sharp peak for $N = 20$ and 28 (GSE), a kink for $N = 18, 22, \ldots, 34$ (GUE), and a smoother connection for $N = 16, 24, 32$ (GOE), for $\beta = 0$. For $\beta = 1$ the feature is preserved,

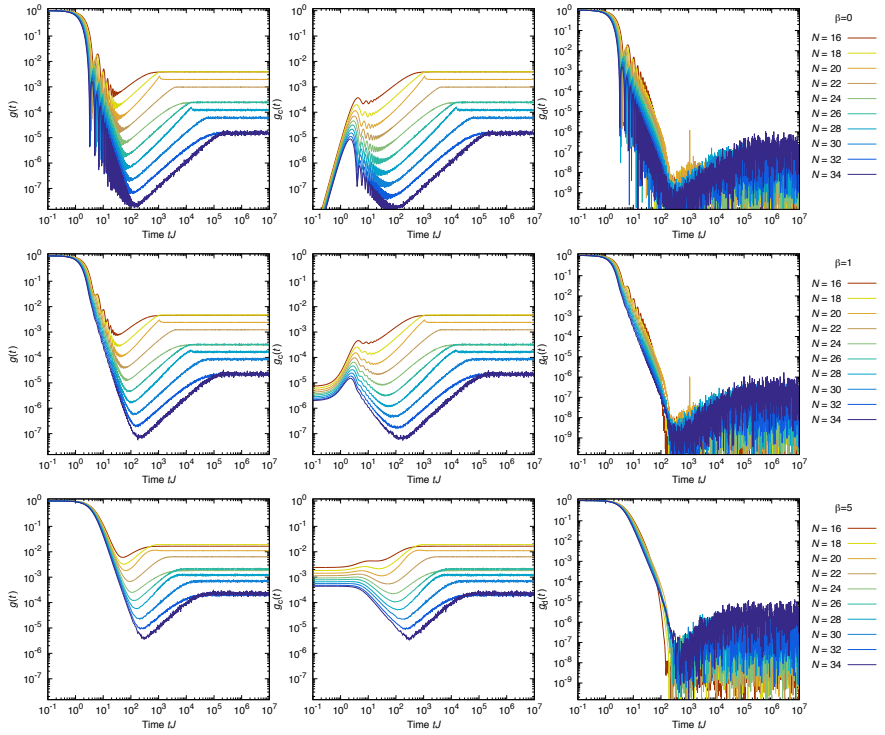


Figure A.2: Plots of $g(t)$, $g_c(t)$, and $g_d(t)$ for $N = 16, 18, \ldots, 34$ and $\beta = 0, 1, 5$, from top to bottom. The noisy part of the curves for g_d are due to the finite number of samples. We expect the true disorder average to continue decreasing rapidly.

while for $\beta = 5$ the peak is broadened and the kink is less visible. However, the plateau heights for $N \not\equiv 0 \pmod 8$ (GUE and GSE) cases appear shifted up compared to those for $N \equiv 0 \pmod 8$ (GOE) cases, for all values of β, and the plateau heights for $N = 18, 26, 34$ are higher than those for $N = 16, 24, 32$ for $\beta > 0$. All of this is consistent with the RMT interpretation, symmetry considerations and smoothing due to unfolding effects.

A.8.2 Dip time t_d, plateau time t_p and plateau height

Intuitively, the dip time can be determined by finding the minimum value of g. However, with finite statistics, the error is large because of the non-self-averaging nature of $g(t)$ past the dip. Therefore, we estimated the error bar as follows. Firstly we found the minimum value $g_{\min}$. Then, the lower and upper limits of the error bar are estimated as the smallest and largest t which give $g(t) < g_{\min} \times 1.04$.

We can fit $t_{\rm dip}$ with an exponential function of N $t_0 e^{\kappa_d N}$. κ_d does not exhibit clear dependence on β from our data (although we expect a weak dependence theoretically). The error bars are large but the results for larger N are consistent with the estimate in Section 2.6. A power-law fit ($t_{\rm d} \sim t_0' N^{\alpha_d}$) cannot be ruled out from our data up to $N = 34$. Again, the available N values are insufficient for a conclusive analysis here.

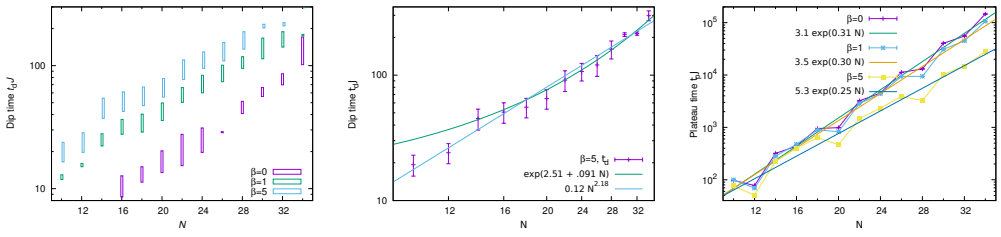


Figure A.3: Left: The dip time $t_{\rm d}$ against N, for $\beta = 0, 1, 5$. The lower and upper limits of the error bar indicate the range of data points with $g(t) < 1.04\ g_{\min}$. Middle: Comparison of fits of the SYK $t_{\rm d}$ with exponential and power-law functions of N for $\beta = 5$. Right: Plot of the plateau time t_p against N, for $\beta = 0, 1, 5$.

As discussed in the main text, the function $g(t)$ reaches a plateau at exponentially late time. Numerically, we find that the height agrees with the expectation $Z(2\beta)/Z(\beta)^2$ when we take an average with sufficiently many samples. The plateau t_p is defined by fitting the ramp by a power-law of the time (linear function

in the log-log plot) and the plateau by a constant, and finding the crossing point of the two lines. We choose the starting point of the fitting range for the ramp as $t_s = 5\, t_d$ if $g(5\, t_d) < 0.7\, g_p$, otherwise we use the time at which $g(t_s) = 0.4\, g_p$. The end of the fitting range is the time at which $g(t_e) = 0.7\, g_p$, and we fit $\log(g(t))$ by a linear fit and find the time at which the line reaches $\log t_p$. In the right panel of Fig. A.3 we plot t_p against N.

As explained in Section 2.6, we expect $t_p \sim \text{const.} \exp(S(2\beta))$. Also, as explained in Section 2.4, the expression for entropy at low temperature is

$$S(2\beta) = (0.23 + 0.198/\beta + \cdots)N + \cdots . \tag{A.52}$$

At $\beta = 5$ the coefficient of N is 0.27 up to $O(\beta^{-2})$ corrections, which is close to our numerical result 0.249 ± 0.014.

As we have seen $t_p \sim \mathrm{e}^{\kappa_p N}$ and $t_d \sim \mathrm{e}^{\kappa_d N}$, where $\kappa_d < \kappa_p = S(2\beta)$. Hence $\log(t_p/t_d)/N \sim \kappa_p - \kappa_d$ should be constant up to $1/N$. We observe that $\kappa_p - \kappa_d > 0$. Therefore, the length of the ramp seems to increase exponentially in N, consistent with Section 2.6. Of course our values of N are not large enough to make definitive statements.

Theoretically the height of plateau of $g(t)$ is $g_p(\beta) = Z(2\beta)/Z(\beta)^2$, (2.5), unless there is degeneracy in the eigenvalues of the model Hamiltonian. In the SYK model, as has been discussed in Secs. 2.3.1 and 2.7, all eigenvalues are doubly degenerate when $N \bmod 8 = 2, 4$ or 6. Therefore we expect $g_p(\beta) = 2Z(2\beta)/Z(\beta)^2$. For $\beta = 0$ this equals the inverse of $Z(\beta = 0) = 2^{N/2}$. For $N \bmod 8 = 0$, on the other hand, we do not expect eigenvalue degeneracy and

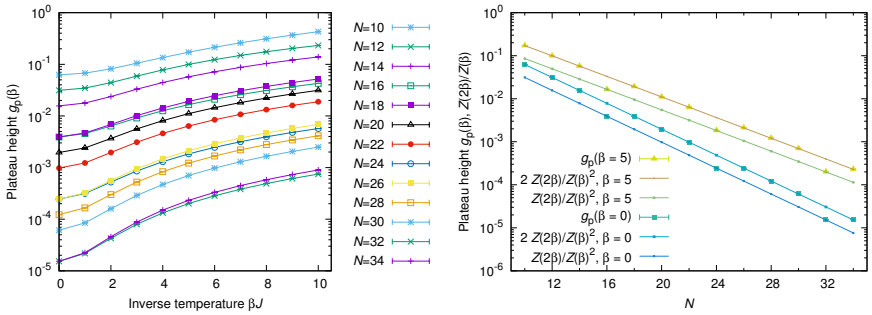


Figure A.4: Left: Plot of the plateau height against β for $N = 10, 12, \ldots, 34$. Right: Plot of the plateau height and $Z(2\beta)/Z(\beta)^2$ for $\beta = 0, 5$ against N. A clear mod–8 pattern can be seen. For $N \equiv 0 \pmod 8$, $g_c(\beta)$ equals $Z(2\beta)/Z(\beta)^2$, which for $\beta = 0$ equals $1/Z(\beta = 0) = 2^{-N/2}$, otherwise $g_c(\beta) = 2Z(2\beta)/Z(\beta)^2$ due to the degeneracy in the eigenvalue of the SYK Hamiltonian.

thus expect $g_p(\beta) = Z(2\beta)/Z(\beta)^2$. We can see nice agreement in Fig. A.4.

A.8.3 Comparison of factorized and unfactorized quantities

As explained in Section 2.3, there are two options for defining the spectral form factor. Namely, the factorized, or annealed, quantities (2.10), (2.11), and (2.12), and the unfactorized, or quenched, versions where one averages over J after dividing by $Z(\beta)^2$. These two choices agree when the quantity of interest is self-averaging (up to order $1/N^q$). Therefore, g and g_u must agree at early time. Numerically we find they agree at large N for all time. $g_c(t)$ is not self averaging at early time and so differs from $g_{uc}(t)$ there.

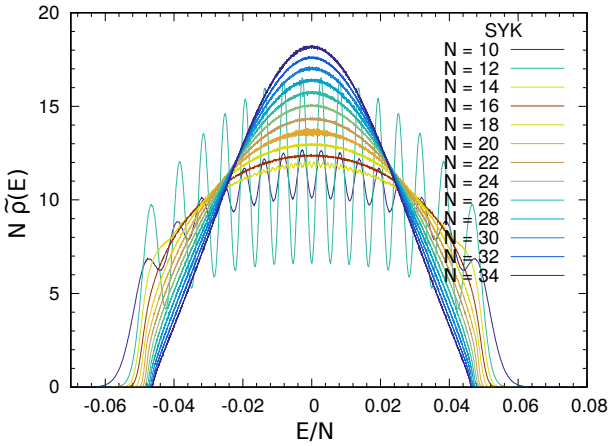


Figure A.5: Normalized density of states $\tilde{\rho}(E)$ for the SYK model with $N = 10, 12, \ldots, 34$. The bin width is $10^{-3}J$. Notice that the energy is measured in units of NJ. The numbers of samples are 21600000 ($N = 10$), 10800000 ($N = 12$), 5400000 ($N = 14$), 1200000 ($N = 16$), 600 000 ($N = 18$), 240 000 ($N = 20$), 120 000 ($N = 22$), 48 000 ($N = 24$), 10 000 ($N = 26$), 3 000 ($N = 28$), 914 ($N = 30$), 516 ($N = 32$), 90 ($N = 34$).

A.8.4 Density of states $\rho(E)$

In Fig. A.5 we plot the normalized density of states $\tilde{\rho}(E)$, averaging the spectrum obtained by diagonalizing the Hamiltonian (2.6) for many disorder parameters. Almost periodic oscillations due to level repulsion are clearly observed for small values of N. For large N and fixed q, the distribution will converge in e.g. an L^2 norm sense to a Gaussian [62], with width $E \sim \sqrt{N}$. However, the small tails of $\tilde{\rho}$ for energies of order $E \sim N$ will not be described by a Gaussian, and will contain an exponentially large number of states.

Bibliography

[1] A. Kitaev, *A simple model of quantum holography, talks at UCSB*, .

[2] J. Maldacena and D. Stanford, *Remarks on the Sachdev-Ye-Kitaev model*, *Phys. Rev.* **D94** (2016), no. 10 106002, [arXiv:1604.0781].

[3] J. Maldacena, D. Stanford, and Z. Yang, *Conformal symmetry and its breaking in two dimensional Nearly Anti-de-Sitter space*, arXiv:1606.0185.

[4] S. Sachdev and J.-w. Ye, *Gapless spin fluid ground state in a random, quantum Heisenberg magnet*, *Phys. Rev. Lett.* **70** (1993) 3339, [cond-mat/9212030].

[5] J. Maldacena, S. H. Shenker, and D. Stanford, *A bound on chaos*, *JHEP* **08** (2016) 106, [arXiv:1503.0140].

[6] J. S. Cotler, G. Gur-Ari, M. Hanada, J. Polchinski, P. Saad, S. H. Shenker, D. Stanford, A. Streicher, and M. Tezuka, *Black holes and random matrices*, *Journal of High Energy Physics* **2017** (May, 2017) 118.

[7] D. A. Roberts, D. Stanford, and A. Streicher, *Operator growth in the SYK model*, *JHEP* **06** (2018) 122, [arXiv:1802.0263].

[8] X.-L. Qi and A. Streicher, *Quantum Epidemiology: Operator Growth, Thermal Effects, and SYK*, arXiv:1810.1195.

[9] J. M. Maldacena, *Eternal black holes in anti-de Sitter*, *JHEP* **04** (2003) 021, [hep-th/0106112].

[10] L. Dyson, J. Lindesay, and L. Susskind, *Is there really a de Sitter/CFT duality?*, *JHEP* **08** (2002) 045, [hep-th/0202163].

[11] G. T. Horowitz and V. E. Hubeny, *Quasinormal modes of AdS black holes and the approach to thermal equilibrium*, *Phys. Rev.* **D62** (2000) 024027, [hep-th/9909056].

[12] K. Papadodimas and S. Raju, *Local Operators in the Eternal Black Hole*, *Phys. Rev. Lett.* **115** (2015), no. 21 211601, [arXiv:1502.0669].

[13] L. Dyson, M. Kleban, and L. Susskind, *Disturbing implications of a cosmological constant*, *JHEP* **10** (2002) 011, [hep-th/0208013].

[14] A. Kitaev, "A simple model of quantum holography." http://online.kitp.ucsb.edu/online/entangled15/kitaev/,http://online.kitp.ucsb.edu/online/entangled15/kitaev2/. Talks at KITP, April 7, 2015 and May 27, 2015.

[15] E. Dyer and G. Gur-Ari, *2D CFT Partition Functions at Late Times*, arXiv:1611.0459.

[16] J. L. F. Barbon and E. Rabinovici, *Very long time scales and black hole thermal equilibrium*, *JHEP* **11** (2003) 047, [hep-th/0308063].

[17] A. L. Fitzpatrick, J. Kaplan, D. Li, and J. Wang, *On information loss in AdS_3/CFT_2*, *JHEP* **05** (2016) 109, [arXiv:1603.0892].

[18] A. L. Fitzpatrick and J. Kaplan, *On the Late-Time Behavior of Virasoro Blocks and a Classification of Semiclassical Saddles*, arXiv:1609.0715.

[19] A. Kitaev. http://online.kitp.ucsb.edu/online/joint98/kitaev/. KITP seminar, Feb. 12, 2015.

[20] J. Polchinski and V. Rosenhaus, *The Spectrum in the Sachdev-Ye-Kitaev Model*, *JHEP* **04** (2016) 001, [arXiv:1601.0676].

[21] S. H. Shenker and D. Stanford, *Black holes and the butterfly effect*, *JHEP* **03** (2014) 067, [arXiv:1306.0622].

[22] A. Kitaev. https://www.youtube.com/watch?v=0Q9qN8j7EZI. Talk given at the Fundamental Physics Prize Symposium, Nov. 10, 2014.

[23] S. H. Shenker and D. Stanford, *Stringy effects in scrambling*, *JHEP* **05** (2015) 132, [arXiv:1412.6087].

[24] A. Almheiri and J. Polchinski, *Models of AdS_2 backreaction and holography*, *JHEP* **11** (2015) 014, [arXiv:1402.6334].

[25] K. Jensen, *Chaos in AdS_2 Holography*, *Phys. Rev. Lett.* **117** (2016), no. 11 111601, [arXiv:1605.0609].

[26] J. Engelsy, T. G. Mertens, and H. Verlinde, *An investigation of AdS_2 backreaction and holography*, *JHEP* **07** (2016) 139, [arXiv:1606.0343].

[27] Y. Gu, X.-L. Qi, and D. Stanford, *Local criticality, diffusion and chaos in generalized Sachdev-Ye-Kitaev models*, *JHEP* **05** (2017) 125, [arXiv:1609.0783].

[28] M. Berkooz, P. Narayan, M. Rozali, and J. Simón, *Higher Dimensional Generalizations of the SYK Model*, arXiv:1610.0242.

[29] W. Fu, D. Gaiotto, J. Maldacena, and S. Sachdev, *Supersymmetric SYK models*, arXiv:1610.0891.

[30] D. J. Gross and V. Rosenhaus, *A Generalization of Sachdev-Ye-Kitaev*, *JHEP* **02** (2017) 093, [arXiv:1610.0156].

[31] A. Jevicki, K. Suzuki, and J. Yoon, *Bi-Local Holography in the SYK Model*, *JHEP* **07** (2016) 007, [arXiv:1603.0624].

[32] A. Almheiri and B. Kang, *Conformal Symmetry Breaking and Thermodynamics of Near-Extremal Black Holes*, *JHEP* **10** (2016) 052, [arXiv:1606.0410].

[33] D. Bagrets, A. Altland, and A. Kamenev, *Sachdev-Ye-Kitaev model as Liouville quantum mechanics*, *Nucl. Phys.* **B911** (2016) 191–205, [arXiv:1607.0069].

[34] M. Cvetic and I. Papadimitriou, *AdS_2 holographic dictionary*, *JHEP* **12** (2016) 008, [arXiv:1608.0701].

[35] W. Fu and S. Sachdev, *Numerical study of fermion and boson models with infinite-range random interactions*, *Phys. Rev.* **B94** (2016), no. 3 035135, [arXiv:1603.0524].

[36] Y.-Z. You, A. W. W. Ludwig, and C. Xu, *Sachdev-Ye-Kitaev Model and Thermalization on the Boundary of Many-Body Localized Fermionic Symmetry Protected Topological States*, arXiv:1602.0696.

[37] A. M. García-García and J. J. M. Verbaarschot, *Spectral and thermodynamic properties of the Sachdev-Ye-Kitaev model*, arXiv:1610.0381.

[38] V. Balasubramanian, M. Berkooz, S. F. Ross, and J. Simon, *Black Holes, Entanglement and Random Matrices*, *Class. Quant. Grav.* **31** (2014) 185009, [arXiv:1404.6198].

[39] D. A. Roberts and B. Yoshida, *Chaos and complexity by design*, *JHEP* **04** (2017) 121, [arXiv:1610.0490].

[40] M. L. Mehta, *Random matrices*, vol. 142. Academic press, 2004.

[41] C. Teitelboim, *Gravitation and Hamiltonian Structure in Two Space-Time Dimensions*, *Phys. Lett.* **B126** (1983) 41–45.

[42] R. Jackiw, *Lower Dimensional Gravity*, *Nucl. Phys.* **B252** (1985) 343–356.

[43] L. Fidkowski and A. Kitaev, *Topological phases of fermions in one dimension*, *Phys. Rev. B* **83** (Feb, 2011) 075103.

[44] F. J. Dyson, *Statistical theory of the energy levels of complex systems. I*, *J. Math. Phys.* **3** (1962) 140–156.

[45] T. Guhr, A. Muller-Groeling, and H. A. Weidenmuller, *Random matrix theories in quantum physics: Common concepts*, *Phys. Rept.* **299** (1998) 189–425, [cond-mat/9707301].

[46] E. Wigner, *Proceedings of the conference on neutron physics by time-of-flight, gatlinburg, tennessee, 1956*, .

[47] O. Bohigas and M.-J. Giannoni, *Chaotic motion and random matrix theories*, in *Mathematical and computational methods in nuclear physics*, pp. 1–99. Springer, 1984.

[48] M. L. Mehta, *On the statistical properties of the level-spacings in nuclear spectra*, *Nuclear Physics* **18** (1960) 395–419.

[49] M. Gaudin, *Sur la loi limite de l'espacement des valeurs propres d'une matrice ale´ atoire*, *Nuclear Physics* **25** (1961) 447–458.

[50] S. Sachdev, *Bekenstein-Hawking Entropy and Strange Metals*, *Phys. Rev.* **X5** (2015), no. 4 041025, [arXiv:1506.0511].

[51] O. Parcollet and A. Georges, *Non-fermi-liquid regime of a doped mott insulator*, *Physical Review B* **59** (1999) 5341.

[52] A. Georges, O. Parcollet, and S. Sachdev, *Quantum fluctuations of a nearly critical heisenberg spin glass*, *Phys. Rev. B* **63** (Mar, 2001) 134406.

[53] J. Polchinski and A. Streicher. Unpublished.

[54] F. J. Dyson, *Statistical theory of the energy levels of complex systems. III*, *Journal of Mathematical Physics* **3** (1962), no. 1 166–175.

[55] E. Brézin and S. Hikami, *Spectral form factor in a random matrix theory*, *Physical Review E* **55** (1997) 4067–4083, [cond-mat/9608116].

[56] E. Brezin and S. Hikami, *Extension of level-spacing universality*, *Phys. Rev.* **E56** (1997) 264–269.

[57] B. Altshuler and B. Shklovskii, *Repulsion of energy levels and conductivity of small metal samples*, *Sov. Phys. JETP* **64** (1986), no. 1 127–135.

[58] E. Brezin and A. Zee, *Universality of the correlations between eigenvalues of large random matrices*, *Nucl. Phys.* **B402** (1993) 613–627.

[59] A. Andreev and B. Altshuler, *Spectral statistics beyond random matrix theory*, *Physical review letters* **75** (1995) 902.

[60] A. Kamenev and M. Mézard, *Wigner-Dyson statistics from the replica method*, *Journal of Physics A Mathematical General* **32** (June, 1999) 4373–4388, [cond-mat/9901110].

[61] D. Stanford and E. Witten, *Fermionic Localization of the Schwarzian Theory*, *JHEP* **10** (2017) 008, [arXiv:1703.0461].

[62] L. Erdős and D. Schröder, *Phase transition in the density of states of quantum spin glasses*, *Mathematical Physics, Analysis and Geometry* **17** (2014) 9164, [arXiv:1407.1552].

[63] Y. Sekino and L. Susskind, *Fast Scramblers*, *JHEP* **10** (2008) 065, [arXiv:0808.2096].

[64] P. Hayden and J. Preskill, *Black holes as mirrors: Quantum information in random subsystems*, *JHEP* **09** (2007) 120, [arXiv:0708.4025].

[65] N. Lashkari, D. Stanford, M. Hastings, T. Osborne, and P. Hayden, *Towards the Fast Scrambling Conjecture*, *JHEP* **04** (2013) 022, [arXiv:1111.6580].

[66] A. Kitaev. Private communication.

[67] J. Deutsch, *Quantum statistical mechanics in a closed system*, *Physical Review A* **43** (1991), no. 4 2046.

[68] M. Srednicki, *Chaos and quantum thermalization*, *Physical Review E* **50** (1994), no. 2 888.

[69] R. Prange, *The spectral form factor is not self-averaging*, *Physical Review Letters* **78** (1997) 2280.

[70] E. Witten, *An SYK-Like Model Without Disorder*, arXiv:1610.0975.

[71] R. Emparan, C. V. Johnson, and R. C. Myers, *Surface terms as counterterms in the AdS / CFT correspondence*, *Phys. Rev.* **D60** (1999) 104001, [hep-th/9903238].

[72] O. J. Dias, J. E. Santos, and B. Way, *Localized $AdS_5 \times S^5$ Black Holes*, *Phys. Rev. Lett.* **117** (2016), no. 15 151101.

[73] L. Susskind, *Some speculations about black hole entropy in string theory*, hep-th/9309145.

[74] G. T. Horowitz, *Comments on black holes in string theory*, *Class. Quant. Grav.* **17** (2000) 1107–1116, [hep-th/9910082].

[75] G. T. Horowitz and J. Polchinski, *A Correspondence principle for black holes and strings*, *Phys. Rev.* **D55** (1997) 6189–6197, [hep-th/9612146].

[76] G. Rodgers and A. Bray, *Density of states of a sparse random matrix*, *Physical Review B* **37** (1988), no. 7 3557.

[77] J. P. Keating, N. Linden, and H. J. Wells, *Random matrices and quantum spin chains*, *Markov Processes and Related Fields* **21** (2014) 537–555, [arXiv:1403.1114].

[78] L. Erdős and H.-T. Yau, *Universality of local spectral statistics of random matrices*, *Bulletin of the American Mathematical Society* **49** (2012), no. 3 377–414, [arXiv:1106.4986].

[79] F. D. M. Haldane, *Continuum dynamics of the 1-D Heisenberg antiferromagnet: Identification with the O(3) nonlinear sigma model*, *Physics Letters A* **93** (1983), no. 9 464 – 468.

[80] E. H. Lieb and D. W. Robinson, *The finite group velocity of quantum spin systems*, *Commun. Math. Phys.* **28** (1972) 251–257.

[81] M. B. Hastings, *Locality in quantum systems*, *Quantum Theory from Small to Large Scales* **95** (2010) 171–212.

[82] D. A. Roberts, D. Stanford, and L. Susskind, *Localized shocks*, *JHEP* **03** (2015) 051, [arXiv:1409.8180].

[83] I. L. Aleiner, L. Faoro, and L. B. Ioffe, *Microscopic model of quantum butterfly effect: out-of-time-order correlators and traveling combustion waves*, *Annals Phys.* **375** (2016) 378–406, [arXiv:1609.0125].

[84] D. A. Roberts and B. Swingle, *Lieb-Robinson Bound and the Butterfly Effect in Quantum Field Theories*, *Phys. Rev. Lett.* **117** (2016), no. 9 091602, [arXiv:1603.0929].

[85] D. Chowdhury and B. Swingle, *Onset of many-body chaos in the $O(N)$ model*, *Phys. Rev.* **D96** (2017), no. 6 065005, [arXiv:1703.0254].

[86] A. A. Patel, D. Chowdhury, S. Sachdev, and B. Swingle, *Quantum butterfly effect in weakly interacting diffusive metals*, *Phys. Rev.* **X7** (2017), no. 3 031047, [arXiv:1703.0735].

[87] Y. Werman, S. A. Kivelson, and E. Berg, *Quantum chaos in an electron-phonon bad metal*, arXiv:1705.0789.

[88] C. von Keyserlingk, T. Rakovszky, F. Pollmann, and S. Sondhi, *Operator hydrodynamics, OTOCs, and entanglement growth in systems without conservation laws*, arXiv:1705.0891.

[89] A. Nahum, S. Vijay, and J. Haah, *Operator Spreading in Random Unitary Circuits*, arXiv:1705.0897.

[90] S. Xu and B. Swingle, *Accessing scrambling using matrix product operators*, arXiv:1802.0080.

[91] C. Dankert, R. Cleve, J. Emerson, and E. Livine, *Exact and approximate unitary 2-designs and their application to fidelity estimation*, *Physical Review A* **80** (2009), no. 1 012304, [quant-ph/0606161].

[92] W. G. Brown and L. Viola, *Convergence rates for arbitrary statistical moments of random quantum circuits*, *Physical review letters* **104** (2010), no. 25 250501.

[93] W. Brown and O. Fawzi, *Scrambling speed of random quantum circuits*, arXiv:1210.6644.

[94] A. Larkin and Y. Ovchinnikov, *Quasiclassical method in the theory of superconductivity*, *JETP* **28,6** (1969) 1200–1205.

[95] A. Almheiri, D. Marolf, J. Polchinski, D. Stanford, and J. Sully, *An Apologia for Firewalls*, *JHEP* **09** (2013) 018, [arXiv:1304.6483].

[96] S. H. Shenker and D. Stanford, *Multiple Shocks*, *JHEP* **12** (2014) 046, [arXiv:1312.3296].

[97] A. Kitaev, *Hidden correlations in the hawking radiation and thermal noise*, 2014. Talk given at the Fundamental Physics Prize Symposium, Nov. 10, 2014.

[98] A. Kitaev, *A simple model of quantum holography*, 2015. Talks at KITP, April 7, 2015 and May 27, 2015.

[99] A. H. Mueller, *Unitarity and the BFKL pomeron*, *Nucl. Phys.* **B437** (1995) 107–126, [hep-ph/9408245].

[100] D. Stanford, *Many-body chaos at weak coupling*, *JHEP* **10** (2016) 009, [arXiv:1512.0768].

[101] G. Tarnopolsky, *On large q expansion in the Sachdev-Ye-Kitaev model*, arXiv:1801.0687.

[102] G. Bentsen, Y. Gu, and A. Lucas, *Fast scrambling on sparse graphs*, arXiv:1805.0821.

[103] P. Hosur and X.-L. Qi, *Characterizing eigenstate thermalization via measures in the Fock space of operators*, *Phys. Rev.* **93** (Apr., 2016) 042138, [arXiv:1507.0400].

[104] S. Xu and B. Swingle, *Locality, Quantum Fluctuations, and Scrambling*, arXiv:1805.0537.

[105] X. Chen and T. Zhou, *Operator scrambling and quantum chaos*, arXiv:1804.0865.

[106] A. Kitaev and S. J. Suh, *The soft mode in the sachdev-ye-kitaev model and its gravity dual*, *Journal of High Energy Physics* **2018** (2018), no. 5 183.

[107] D. Stanford, *Many-body chaos at weak coupling*, *Journal of High Energy Physics* **2016** (2016), no. 10 9.

[108] A. Lucas, *Operator size at finite temperature and planckian bounds on quantum dynamics*, *arXiv preprint arXiv:1809.07769* (2018).

[109] J. Maldacena and X.-L. Qi, *Eternal traversable wormhole*, arXiv:1804.0049.

[110] Y. Gu, A. Lucas, and X.-L. Qi, *Spread of entanglement in a Sachdev-Ye-Kitaev chain*, *JHEP* **09** (2017) 120, [arXiv:1708.0087].

[111] I. Peschel, *Calculation of reduced density matrices from correlation functions*, *Journal of Physics A Mathematical General* **36** (Apr., 2003) L205–L208, [cond-mat/0212631].

[112] A. Eberlein, V. Kasper, S. Sachdev, and J. Steinberg, *Quantum quench of the sachdev-ye-kitaev model*, *Physical Review B* **96** (2017), no. 20 205123.

[113] A. Streicher, *The Large-q SYK Four-Point Function for All Coupling*, . To appear late Nov.

[114] D. Stanford and L. Susskind, *Complexity and Shock Wave Geometries*, *Phys. Rev.* **D90** (2014), no. 12 126007, [arXiv:1406.2678].

[115] D. A. Roberts and D. Stanford, *Two-dimensional conformal field theory and the butterfly effect*, *Phys. Rev. Lett.* **115** (2015), no. 13 131603, [arXiv:1412.5123].

[116] A. R. Brown, H. Gharibyan, A. Streicher, L. Susskind, L. Thorlacius, and Y. Zhao, *Falling Toward Charged Black Holes*, arXiv:1804.0415.

[117] P. Saad, S. H. Shenker, and D. Stanford, *A semiclassical ramp in SYK and in gravity*, arXiv:1806.0684.

[118] P. Saad, S. H. Shenker, and D. Stanford, *JT gravity as a matrix integral*, arXiv:1903.1111.

[119] H. W. Lin, J. Maldacena, and Y. Zhao, *Symmetries Near the Horizon*, arXiv:1904.1282.

[120] M. E. Ismail, D. Stanton, and G. Viennot, *The combinatorics of q-hermite polynomials and the askey wilson integral*, *European Journal of Combinatorics* **8** (1987), no. 4 379–392.

[121] J. M. Magan, *Random free fermions: An analytical example of eigenstate thermalization*, *Phys. Rev. Lett.* **116** (2016), no. 3 030401, [arXiv:1508.0533].

[122] D. Anninos, T. Anous, and F. Denef, *Disordered Quivers and Cold Horizons*, arXiv:1603.0045.

[123] C. Itzykson and J. B. Zuber, *The Planar Approximation. 2.*, *J. Math. Phys.* **21** (1980) 411.